

OXFORD

ARTIFICIAL INTELLIGENCE AND HUMAN RIGHTS

edited by

Alberto Quintavalla
Jeroen Temperman

Artificial Intelligence and Human Rights

Artificial Intelligence and Human Rights

Edited by

ALBERTO QUINTAVALLA
JEROEN TEMPERMAN

OXFORD
UNIVERSITY PRESS



Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© The Several Contributors 2023

The moral rights of the authors have been asserted

First Edition published in 2023

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Public sector information reproduced under Open Government Licence v3.0
(<http://www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm>)

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2023936581

ISBN 978–0–19–288248–6

DOI: 10.1093/oso/9780192882486.001.0001

Printed and bound in the UK by
TJ Books Limited

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

Acknowledgements

Early versions of roughly half of this volume's chapters were presented on 28 October 2021 at a hybrid conference entitled 'Artificial Intelligence & Human Rights: Friend or Foe?' The expert meeting, organised by the editors, fell right in between two COVID-19 lockdown periods and was held at Erasmus University Rotterdam. The editors would wish to thank the conference's sponsors for their generous funding and support, the Netherlands Network for Human Rights Research, the Jean Monnet Centre of Excellence on Digital Governance, and the Research Office at Erasmus School of Law. Thanks to Hannah Driesens and Jitske de Vries for their invaluable help during the editing process and for completing the law tables.

Contents

<i>Table of International Law</i>	xi
<i>Table of Domestic Law</i>	xv
<i>Table of International Cases</i>	xxi
<i>Table of Domestic Cases</i>	xxv
<i>Abbreviations</i>	xxix
<i>About the Contributors</i>	xxxv

PART I AI-BASED HUMAN RIGHTS VIOLATIONS: LEGAL AND TECHNICAL BACKGROUND

1. Introduction	3
<i>Alberto Quintavalla and Jeroen Temperman</i>	
2. AI Life Cycle and Human Rights: Risks and Remedies	16
<i>Martina Šmuclerová, Luboš Král, and Jan Drchal</i>	

PART II ARTIFICIAL INTELLIGENCE AND ASSORTED FIRST GENERATION CIVIL AND POLITICAL RIGHTS

3. Artificial Intelligence and the Right to Liberty and Security	45
<i>Valentina Golenova</i>	
4. Artificial Intelligence and Religious Freedom	61
<i>Jeroen Temperman</i>	
5. Artificial Intelligence and Freedom of Expression	76
<i>Giovanni De Gregorio and Pietro Dunn</i>	
6. Artificial Intelligence and Freedom of Assembly	91
<i>Margaret Warthon</i>	
7. Artificial Intelligence and the Right to Property: The Human Rights Dimension of Intellectual Property	104
<i>Letizia Tomada and Raphaële Xenidis</i>	

PART III ARTIFICIAL INTELLIGENCE AND PRIVACY

8. Artificial Intelligence and the Right to Privacy	121
<i>Alessia Zornetta and Ignacio Cofone</i>	

9. The Rights to Privacy and Data Protection and Facial Recognition Technology in the Global North <i>Natalia Menéndez González</i>	136
10. Privacy, Political Participation, and Dissent: Facial Recognition Technologies and the Risk of Digital Authoritarianism in the Global South <i>Malcolm Katrak and Ishita Chakrabarty</i>	150
11. The Production of and Control over Data in the AI-Era: The Two Failing Approaches to Privacy Protection <i>Bart van der Sloot</i>	162
12. Artificial Intelligence, the Public Space, and the Right to Be Ignored <i>Andrea Pin</i>	177

PART IV ARTIFICIAL INTELLIGENCE AND NON-DISCRIMINATION

13. Artificial Intelligence and Racial Discrimination <i>Louis Koen and Kgomoitso Mufamadi</i>	195
14. Artificial Intelligence and Gender-Based Discrimination <i>Fabian Lütz</i>	207
15. Artificial Intelligence and LGBTQ+ Rights <i>Masuma Shahid</i>	222
16. Artificial Intelligence and Women's Rights: Deepfake Technology <i>Marília Papaléo Gagliardi</i>	235
17. Artificial Intelligence and Disability Rights <i>Antonella Zarra, Silvia Favalli, and Matilde Ceron</i>	248

PART V ARTIFICIAL INTELLIGENCE AND FAIR PROCEDURE

18. Artificial Intelligence and Fair Trial Rights <i>Helga Molbæk-Steenig and Alexandre Quemy</i>	265
19. Artificial Intelligence and Data Analytics: A Recipe for Human Rights Violations <i>Migle Laukyte</i>	281
20. Artificial Intelligence and the Right to an Effective Remedy <i>Sarah de Heer</i>	294

PART VI ARTIFICIAL INTELLIGENCE AND ASYLUM		
21. Artificial Intelligence Technologies and the Right to Seek and Enjoy Asylum: An Overview <i>Raimy Reyes</i>		311
22. Artificial Intelligence Screening and the Right to Asylum <i>Dhruv Somayajula</i>		327
PART VII ARTIFICIAL INTELLIGENCE AND SECOND GENERATION RIGHTS		
23. Artificial Intelligence and the Right to Food <i>Adekemi Omotubora</i>		343
24. Artificial Intelligence and the Right to Housing <i>Caroline Compton and Jessie Hohmann</i>		355
25. Artificial Intelligence and Human Rights at Work <i>Joe Atkinson and Philippa Collins</i>		371
26. Artificial Intelligence and the Right to Health <i>Enrique Santamaría Echeverría</i>		386
PART VIII ARTIFICIAL INTELLIGENCE AND THIRD GENERATION RIGHTS		
27. Artificial Intelligence and Consumer Protection Rights <i>Shu Li, Béatrice Schütte, and Lotta Majewski</i>		405
28. Artificial Intelligence and the Right to a Healthy Environment <i>Alberto Quintavalla</i>		425
PART IX ARTIFICIAL INTELLIGENCE AND HUMAN RIGHTS: REFLECTIONS		
29. Artificial Intelligence and Human Rights: Understanding and Governing Common Risks and Benefits <i>Kostina Prifti, Alberto Quintavalla, and Jeroen Temperman</i>		441
30. Human Rights, Legal Personality, and Artificial Intelligence: What Can Epistemology and Moral Philosophy Teach Law? <i>Klaus Heine</i>		458
31. Robot Rights/Human Responsibility <i>David Gunkel</i>		471

x CONTENTS

32. The Limits of AI Decision-Making: Are There Decisions Artificial Intelligence Should Not Make? <i>Florian Gamper</i>	484
33. Smart Cities, Artificial Intelligence, and Public Law: An Unchained Melody <i>Sofia Ranchordás</i>	501
34. Putting Private Sector Responsibility in the Mix: A Business and Human Rights Approach to Artificial Intelligence <i>Isabel Ebert and Lisa Hsin</i>	517
35. Artificial Intelligence Human Rights Impact Assessment <i>Alessandro Ortalda and Paul De Hert</i>	531
36. Real-Life Experimentation with Artificial Intelligence <i>Elizaveta Gromova and Evert Stamhuis</i>	551
PART X CONCLUSION	
37. Conclusion <i>Alberto Quintavalla and Jeroen Temperman</i>	569
<i>Bibliography</i>	571
<i>Index</i>	613

Table of International Law

AFRICAN UNION

<i>African Charter on Human and Peoples' Rights</i>	45, 93, 105, 154–55, 159–61, 266, 284, 295, 296–97, 386–87, 432
<i>African Charter on the Rights and Welfare of the Child</i>	154–55
<i>Protocol to the African Charter on Human and Peoples' Rights on the Rights of Women in Africa</i>	213

COUNCIL OF EUROPE

<i>Additional Protocol [No 1] to the Convention for the Protection of Human Rights and Fundamental Freedoms</i>	105, 107–8
<i>Convention 108+ for the Protection of Individuals with Regard to the Processing of Personal Data</i>	124–25, 128, 130–31, 137, 139–40, 147–48, 352, 535
<i>[European] Convention for the Protection of Human Rights and Fundamental Freedoms</i>	45, 58, 76, 77, 93, 107–8, 167, 213, 219–20, 265–66, 377–78, 379–80, 382, 386–87, 416, 532, 552–53, 557, 558, 559, 560, 562, 565
Article 6	112, 265, 266
Article 8	112, 121–22, 169
Article 9	210, 211–12
Article 10	111–12
Article 11	99
Article 13	284, 295
Article 14	19, 112, 213, 284, 416–17
<i>European Social Charter</i>	110–11
<i>Protocol No 9 to the Convention for the Protection of Human Rights and Fundamental Freedoms</i>	169–70
<i>Protocol No 11 to the Convention for the Protection of Human Rights and Fundamental Freedoms</i>	169–70
<i>Protocol No 12 to the Convention for the Protection of Human Rights and Fundamental Freedoms</i>	213
<i>Protocol No 14 to the Convention for the Protection of Human Rights and Fundamental Freedoms</i>	169–70

EUROPEAN UNION

<i>AI Act (draft) (Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts COM/2021/206 final)</i>	34, 56–57, 72, 117–18, 211, 218–20, 255, 258–59, 261, 271, 323, 355–56, 381–83, 406–7, 455–56, 533, 536, 537–38, 541, 545, 552, 554–55, 556–57, 558, 559, 560–61, 566
--	--

<i>Charter of Fundamental Rights of the European Union</i>	70–71, 121–22, 295, 382, 386–87, 544, 556–57
Article 7	121–22, 167, 352
Article 8	122–23, 352
Article 1177, 111–12
Article 17105, 106–7
Article 34	358
Article 37	432
Article 47	265, 295
<i>Digital Services Act</i>	88, 90, 228–29, 421–22, 523, 529–30
<i>Digital Markets Act</i>	228–29, 421–23
<i>European Charter of Patients' Rights</i>289–90
<i>GDPR (European Parliament and Council Regulation 2016/679 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data [2016] OJ L 119/1)</i>	104, 122–23, 124, 126, 133–34, 137, 139, 141–43, 147–48, 168, 228–29, 282–83, 306–7, 382–83, 411–12, 420–21, 449–50, 529–30, 532–33, 538–39, 540, 541, 542, 549–50, 566
Article 2	123–24, 536–37
Article 4	123–24, 139, 141, 171–72, 288, 306, 352
Article 5	124–25, 128, 129–31, 133, 134–35, 166, 168, 306, 352, 396
Article 6125, 128, 134
Article 7172
Article 8147–48
Article 9139–40, 396–97
Article 13127
Article 14127
Article 15113–14, 127
Article 17130–31
Article 18130–31
Article 21419–20
Article 22125–26, 219, 322–23, 337
Article 24133, 134–35
Article 25140, 422
Article 26124
Article 28133
Article 33131
Article 34131
Article 35133–34, 305, 382, 535–36, 538–41, 549
Article 36133–34
Article 37133
Article 77133
Article 82133
Article 83133
Article 89129, 133–34
Article 90133–34
Article 91133–34
Article 92133–34
Article 93133–34

Article 94.....	133–34
Article 95.....	133–34
<i>Treaty on the Functioning of the European Union</i>	122–23, 218–19, 558

ORGANIZATION OF AMERICAN STATES

<i>Additional Protocol to the American Convention on Human Rights in the Area of Economic, Social and Cultural Rights (San Salvador Protocol)</i>	386–87, 432
<i>American Convention on Human Rights</i>	45, 167, 295
Article 8.....	265–66
Article 11.....	155–56, 159–61
<i>Inter-American Convention Against all Forms of Discrimination And Intolerance</i>	213

UNITED NATIONS

<i>Aarhus Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters</i>	295, 432
<i>Convention on the Elimination of All Forms of Discrimination against Women</i>	207–8, 212–13, 216–17, 220–21, 241, 329–30
<i>Convention on the Rights of the Child</i>	148, 329–30, 534–35
<i>Convention on the Rights of Persons with Disabilities</i>	10, 249–50, 255–58, 260, 329–30, 364–65
<i>International Covenant on Civil and Political Rights</i>	167, 228–29, 266, 328–29, 377–78, 475
Article 2.....	203–4, 295, 346–47, 520
Article 6.....	485
Article 9.....	45, 485
Article 14.....	26, 265, 266, 485
Article 17.....	121–22, 159, 241
Article 18.....	63–64
Article 19.....	76
Article 21.....	93
Article 26.....	19, 329–30
<i>International Covenant on Economic, Social and Cultural Rights</i>	110–11, 228–29, 353, 356, 359, 380, 475
Article 2.....	346–48, 357
Article 11.....	343–45, 346–48, 349–50, 356–57, 367
Article 12.....	367, 386–87
Article 15.....	105, 106–7
<i>International Convention on the Elimination of All Forms of Racial Discrimination</i>	198–99, 202, 203–4, 329–30
Article 2.....	200
Article 5.....	200
<i>Convention Relating to the Status of Refugees</i>	314, 318, 328–29
<i>Universal Declaration of Human Rights</i>	3, 64, 93, 167, 207–8, 228–29, 241, 281, 328–29, 344, 356–57, 459, 475, 534–35
Article 8.....	295, 520
Article 10.....	265
Article 12.....	121–22, 241, 352, 412

xiv TABLE OF INTERNATIONAL LAW

Article 14.....	11, 311, 327–28, 329
Article 19.....	76
Article 22.....	345
Article 23.....	345
Article 27.....	105, 345
<i>United Nations Convention on the Law of the Sea</i>	295

OTHER TREATIES

<i>Arab Charter on Human Rights</i> (League of Arab States)	432
<i>European Patent Convention</i>	108–9
<i>TRIPS Agreement</i> (WTO)	105–6, 108

Table of Domestic Law

ALBANIA

- Constitution of Albania (2008) 106–7

ALGERIA

- Constitution of Algeria (2020) 154–55

ANDORRA

- Personal Data Protection Act 2021 122–23

ARGENTINA

- Constitution of Argentina (1853) 155–56
Personal Data Protection Act 2000 122–23

AUSTRALIA

- Equal Opportunity Act [Victoria, Australia] Act No 16 of 2010 205
Modern Slavery Act 2018 521
Privacy Act 1998 130–31
Racial Discrimination Act 1975 204
Sex Discrimination Act 1984 213

BANGLADESH

- Constitution of Bangladesh (1972) 152–53
Digital Security Act 2018 152–53
Information and Communications Technology Act 2006 152–53
Telecommunication Act 2001 152–53
Telegraph Act 1886 152–53

BENIN

- Constitution of Benin (1990) 154–55

BOLIVIA

- Constitution of Bolivia (2009) 155–56

BRAZIL

- Constitution of Brazil (1988) 155–56
Lei Geral de Proteção de Dados Pessoais Lei No 13.709 (2018) 125–26, 535

CANADA

An Act to Modernize Legislative Provisions as regards the Protection of Personal Information [Quebec].....	133
Bill S-211 (An Act to enact the Fighting Against Forced Labour and Child Labour in Supply Chains Act and to amend the Customs Tariff)	521
Directive on Automated Decision-Making (2021).....	213–14
Personal Information Protection and Electronic Documents Act 2000	122–23, 127

CAPE VERDE

Constitution of Cape Verde (1980).....	154–55
--	--------

CHINA

Constitution of China (1982)	152–53
------------------------------------	--------

COLOMBIA

Constitution of Colombia (1991)	155–56
---------------------------------------	--------

DOMINICAN REPUBLIC

Constitution of Dominican Republic (2015).....	155–56
--	--------

ECUADOR

Constitution of Ecuador (2008)	155–56
--------------------------------------	--------

ERITREA

Constitution of Eritrea (1997).....	154–55
-------------------------------------	--------

FRANCE

Law No 2019-222 of 23 March 2019 [on Programming and Reform of the Justice System]	286, 287
--	----------

GERMANY

Act on Corporate Due Diligence in Supply Chains 2021.....	523
Basic Law for the Federal Republic of Germany (1949).....	265

GHANA

Constitution of Ghana (1992).....	154–55
-----------------------------------	--------

INDIA

Personal Data Protection Bill (No 373 of 2019)	129, 356–57
--	-------------

ISRAEL

Protection of Privacy Law (No 5741 of 1981)	122–23
---	--------

JAPAN

- Act on the Protection of Personal Information
 (No 57 of 2003 amended 2020) 122–23, 128

KENYA

- Constitution of Kenya (2010) 154–55

MALAWI

- Constitution of Malawi (1994) 154–55

MEXICO

- Constitution of Mexico (1917) 155–56

MONTENEGRO

- Constitution of Montenegro (2007) 106–7

MOZAMBIQUE

- Constitution of Mozambique (2004) 154–55

NEPAL

- Constitution of Nepal (2015) 152–53

NEW ZEALAND

- Privacy Act 2020 122–23

NICARAGUA

- Constitution of Nicaragua (1987) 155–56

NORWAY

- Act Relating to Enterprises' Transparency and Work on Fundamental Human
 Rights and Decent Working Conditions (Transparency Act) 2021 523

PANAMA

- Constitution of Panama (1972) 155–56

PARAGUAY

- Constitution of Paraguay (1992) 155–56

PERU

- Constitution of Peru (1993) 155–56
 Ley de Protección de Dados Personales (No 29733 of 2011) 129

PORUGAL

Constitution of Portugal (1976)	106–7
---------------------------------------	-------

SENEGAL

Constitution of Senegal (2001)	154–55
--------------------------------------	--------

SOUTH AFRICA

Constitution of South Africa (1996).....	154–55
Promotion of Equality and Prevention of Unfair Discrimination Act (PEPUDA) Act 4 of 2000.....	204–5, 356–57

SOUTH KOREA

Constitution of the Republic of Korea (1948)	152–53
Personal Information Protection Act (2011, amended 2020)	535

SRI LANKA

Personal Data Protection Act 2022 (Act No 9 of 2022)	152–53
Right to Information Act 2016 (Act No 12 of 2016)	153

SWEDEN

Constitution of Sweden (1974)	106–7
-------------------------------------	-------

THAILAND

Personal Data Protection Act 2020.....	152–53
--	--------

UNITED KINGDOM

Equality Act 2010	201–2, 204, 334, 376, 382
Modern Slavery Act 2015.....	521
Online Safety Bill (pending 2022).....	523

UNITED STATES

Automated Employment Decision Tools (Law No 144 of 2021).....	36
Biometric Information Privacy Act 2008.....	137, 139, 141, 143
California Civil Code 1872	288
California Consumer Privacy Act 2018	130–31
California Privacy Rights Act 2020.....	130
Civil Rights Act 1964	213
Constitution of the United States (1789)	77, 87, 122, 265
Dodd-Frank Act 2010 (Wall Street Reform and Consumer Protection Act)	521–22
National AI Initiative Act 2020.....	213–14
New York City Int 1894–2020 (Local Law to Amend the Administrative Code of the City of New York)	208, 218–19
Public Oversight of Surveillance Technology (POST) Act 2018.....	57–58
San Francisco Administrative Code (amended 2022)	224–25, 233–34

URUGUAY

- Ley no 18.331 de Protección de Datos Personales y Acción de
Habeas Data (2008) 125–26

VENEZUELA

- Constitution of Venezuela (1999) 155–56

ZIMBABWE

- Constitution of Zimbabwe (2013) 154–55

Table of International Cases

AFRICAN UNION

African Commission on Human and Peoples' Rights

- Social and Economic Rights Action Center (SERAC) and Center for Economic and Social Rights (CESR) v Nigeria* Communication No 155/96
(27 May 2002). 347, 433
- Zimbabwe Human Rights NGO Forum v Zimbabwe* Communication No 245/2002
(25 May 2006). 203–4

COUNCIL OF EUROPE

European Court of Human Rights (incl. former EcommHR decisions)

- Airey v Ireland* App no 6289/73 (9 October 1979). 559
- Amann v Switzerland* App no 27798/95 (16 February 2000). 122–23
- Amuur v France* App no 19776/92 (25 June 1996) 46–47
- Anheuser-Busch Inc v Portugal* App no 73049/01 (11 January 2007) 105
- Antović and Mirković v Montenegro* App no 70838/13 (28 November 2017) 379
- Aral and Tekin v Turkey* App no 24563/94 (ECommHR, 14 January 1998). 107–8
- Asselbourg and 78 Others and Greenpeace Association-Luxembourg v Luxembourg*
App no 29121/95 (29 June 1999) 169–70
- Bărbulescu v Romania* App no 61496/08 (5 September 2017). 379
- Beyeler v Italy* App no 33202/96 (5 January 2000). 107–8
- Biao v Denmark* App no 38590/10 (24 May 2016) 19
- Big Brother Watch and Others v United Kingdom* App nos 58170/13,
62322/14 and 24960/15 (25 May 2021). 100–1, 213, 447–48
- Broniowski v Poland* App no 31443/96 (28 September 2005) 107–8
- Carson and Others v United Kingdom* App no 42184/05 (16 March 2010). 19
- Catt v United Kingdom* App no 43514/15 (24 January 2019). 99–100
- Church of Scientology of Paris v France* App no 19509/92 (ECommHR,
9 January 1995). 169–70
- Cumhuriyet Vakfi and Others v Turkey* App no 28255/07 (8 October 2013) 100–1
- DH and Others v Czech Republic* App no 57325/00 (13 November 2007) 19, 219–20
- Dhahbi v Italy* App no 17120/09 (8 April 2014). 297–98
- Dima v Romania* App no 58472/00 (26 May 2005). 107–8
- Djavit An v Turkey* App no 20652/92 (20 February 2003). 99
- Dombo Beheer BV v The Netherlands* App no 14448/88 (27 October 1993). 296–97
- Ernst and Others v Belgium* App no 33400/96 (15 July 2003). 296–97
- Ezelin v France* App no 11800/85 (26 April 1991). 99
- Fadeyeva v Russia* App no 55723/00 (9 June 2005). 435–36
- Former King of Greece v Greece* App no 25701/94 (28 November 2002) 107–8
- Forrer-Niedenthal v Germany* App no 47316/99 (20 February 2003) 107–8
- Gaughran v United Kingdom* App no 45245/15 (13 June 2020). 100
- Gülçü v Turkey* App no 17526/10 (6 June 2016) 93

<i>Hadjanastassiou v Greece</i> App no 12945/87 (16 December 1992)	297–98
<i>Handyside v United Kingdom</i> App no 5493/72 (7 December 1976)	563
<i>Hatton v United Kingdom</i> App no 36022/97 (8 July 2003)	433
<i>Hoogendijk v The Netherlands</i> App no 58641/00 (6 January 2005)	219–20
<i>Hugh Jordan v United Kingdom</i> App no 24746/94 (4 April 2001)	219–20
<i>II v Bulgaria</i> App no 44082/98 (9 June 2005)	46–47
<i>Iilan v Turkey</i> App no 22277/93 (27 June 2010)	295
<i>Jersild v Denmark</i> App No 15890/89 (23 September 1994)	80
<i>Klass and Others v Germany</i> App no 5029/71 (6 September 1978)	295
<i>Krone Verlag GmbH & Co KG v Austria (No 3)</i> App no 39069/97 (11 December 2003)	110–11
<i>Kudrevičius and Others v Lithuania</i> App no 37553/05 (15 October 2015)	93, 99
<i>Kyrtatos v Greece</i> App no 41666/98 (22 May 2003)	435–36
<i>Lawlor v United Kingdom</i> App no 12763/87 (14 July 1988)	169–70
<i>LCB v United Kingdom</i> App no 23413/94 (9 June 1998)	559
<i>Lenzing AG v United Kingdom</i> App no 38817/97 (ECommHR, 9 September 1998)	107–8
<i>Makhfi v France</i> App no 59335/00 (19 October 2004)	296–97
<i>Matos and Silva Lda v Portugal</i> App no 15777/89 (16 September 1996)	107–8
<i>McFarlane v Ireland</i> App no 31333/06 (10 September 2010)	295
<i>Megadat.com SRL v Moldova</i> App no 21151/04 (8 April 2004)	296–97
<i>Molla Sali v Greece</i> App no 20452/14 (19 December 2018)	19
<i>Nachova and Others v Bulgaria</i> App nos 43577/98 and 43579/98 (6 July 2005)	112–13
<i>Nideröst-Huber v Switzerland</i> App no 18990/91 (18 February 1997)	296–97
<i>Observer and Guardian v United Kingdom</i> App No 13585/88 (26 November 1991)	80
<i>Öcalan v Turkey</i> App no 46221/99 (12 May 2005)	266
<i>Opuz v Turkey</i> App no 33401/02 (9 June 2009)	219–20
<i>Orsus and Others v Croatia</i> App no 15766/03 (16 March 2010)	219–20
<i>Osman v United Kingdom</i> App no 23452/94 (28 October 1998)	559
<i>Othman (Abu Qatada) v United Kingdom</i> App no 8139/09 (17 January 2012)	316
<i>PG v United Kingdom</i> App no 44787/98 (25 September 2001)	186
<i>PN v Germany</i> App no 74440/17 (16 November 2020)	100
<i>Roman Zakharov v Russia</i> App no 47143/06 (4 December 2015)	122
<i>Rotaru v Romania</i> App no 28341/95 (4 May 2000)	122–23
<i>Ruiz Rivera v Switzerland</i> App no 8300/06 (18 February 2014)	268
<i>Ruiz Torija v Spain</i> App no 18390/91 (9 December 1994)	297–98
<i>S and Marper v United Kingdom</i> App nos 30562/04 and 30566/04 (4 December 2008)	99, 143–44
<i>SAS v France</i> App no 43835/11 (1 July 2014)	561
<i>Schuler-Zgraggen v Switzerland</i> App no 14518/89 (24 June 1993)	296–97
<i>Schwabe and M.G. v Germany</i> App nos 8080/08 and 8577/08 (1 March 2012)	100
<i>Segerstedt-Wiberg and Others v Sweden</i> App no 62332/00 (6 September 2006)	99–100
<i>Smith Kline and French Laboratories Ltd v the Netherlands</i> App no 12633/87 (ECommHR, 4 October 1990)	107–8
<i>Soering v United Kingdom</i> App no 14038/88 (7 July 1989)	314
<i>Storck v Germany</i> App no 61603/00 (16 June 2005)	47
<i>Szabo v Hungary</i> App no 37138/14 (12 January 2016)	100–1
<i>Tauira and Others v France</i> App no 28204/95 (ECommHR, 4 December 1995)	169–70
<i>Tyrer v United Kingdom</i> App no 5856/72 (25 April 1978)	532
<i>Velikova v Bulgaria</i> App no 41488/98 (8 May 2000)	219–20
<i>VK v Russia</i> App no 9139/08 (4 April 2017)	46–47
<i>Weber and Saravia v Germany</i> App no 54934/00 (29 June 2006)	100–1

<i>X and Church of Scientology v Sweden</i> App no 7805/77 (ECommHR, 5 May 1979)	66, 212–13
<i>X and Y v The Netherlands</i> App no 8978/80 (26 March 1985).....	559
<i>Zakharov and Varzhabetyan v Russia</i> App nos 35880/14 and 75926/17 (13 January 2021).....	99–100

European Committee of Social Rights

<i>Centre on Housing Rights and Evictions (COHRE) v France</i> Complaint no 63/2010 (28 June 2011).....	358–59
<i>Centre on Housing Rights and Evictions (COHRE) v Italy</i> Complaint no 58/2009 (25 June 2010).....	358–59
<i>European Federation of National Organisation/s working with the Homeless (FEANTSA) v France</i> Complaint no 39/2009 (5 December 2007).....	358–59
<i>European Federation of National Organisations Working with the Homeless (FEANTSA) v Slovenia</i> Complaint no 53/2008 (8 September 2009)	358–59
<i>European Roma Rights Centre v Greece</i> Complaint no 15/2003 (7 February 2005).....	358–59
<i>European Roma Rights Centre v Portugal</i> Complaint no 61/2010 (30 June 2011).....	358–59

EUROPEAN UNION

European Court of Justice

<i>Case C-101/01 Lindqvist</i> (2003)	78–79
<i>Case C-104/10 Kelly</i> (2011)	219
<i>Case C-109/88 Danfoss</i> (1989)	112–13, 219–20
<i>Case C-131/12 Google Spain SL and Google Inc v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González</i> (2014)	78–79
<i>Case C-169/14 Sánchez Morcillo</i> (2014).....	296–97
<i>Case C-199/11 European Community v Otis NV and Others</i> (2012).....	296–97
<i>Case C-283/05 ASML Netherlands BV</i> (2006)	297–98
<i>Case C-291/12 Michael Schwarz v Stadt Bochum</i> (2014)	122
<i>Case C-274/18 Schuch-Ghannadan</i> (2019)	219–20
<i>Cases C-293/12 and C594/12 (joined cases) Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others</i> (2014)	47, 122
<i>Case C-362/14 Maximillian Schrems v Data Protection Commissioner</i> (2014)	130–31
<i>Case C-393/09 Bezpecnostní Softwarová Asociace</i> (2010)	109–10
<i>Cases C-403/08 and C-429/08 (joined cases) Football Association Premier League and Others</i> (2011)	109–10
<i>Case C-406 SAS Institute</i> (2012)	109–10
<i>Case C-415/10 Meister</i> (2012).....	219
<i>Case C-434/16 Peter Nowak v Data Protection Commissioner</i> (2017)	306
<i>Case C-465/00 Rechnungshof v Österreichischer Rundfunk and Others</i> (2003)	125
<i>Case C-471/11 Banif Plus Bank</i> (2013)	296
<i>Case C-5/08 Infopaq International A/S v Danske Dagblades Forening</i> (2009)	109–10
<i>Case C-524/06 Heinz Huber v Bundesrepublik Deutschland</i> (2008)	296
<i>Case C-55/94 Reinhard Gebhard v Consiglio dell'Ordine degli Avvocati e Procuratori di Milano</i> (1995)	145–46
<i>Case C-582/14 Patrick Beyer v Bundesrepublik Deutschland</i> (2016).....	306
<i>Case C-619/10 Trade Agency Ltd</i> (2012)	297–98
<i>Case C-89/08 European Commission v Ireland</i> (2009)	296

ORGANIZATION OF AMERICAN STATES

Inter-American Court of Human Rights

<i>Artavia Murillo et al (In Vitro Fertilization) v Costa Rica</i> Ser C No 257 (28 November 2012)	155–56
<i>Pacheco Tineo Family v Bolivia</i> Ser C No 272 (25 November 2013)	316
<i>Rights and Guarantees of Children in the Context of Migration and/or in Need of International Protection</i> Advisory Opinion OC-21/14 (19 August 2014)	320
<i>State Obligations Concerning the Change of Name, Gender Identity, and Rights Derived from a Relationship Between Same-Sex Couples</i> Advisory Opinion OC-24/17 (24 November 2017)	231–32
<i>The Environment and Human Rights (State Obligations in Relation to the Environment in the Context of the Protection and Guarantee of the Rights to Life and to Personal Integrity—Interpretation and Scope of Articles 4(1) and 5(1) of the American Convention on Human Rights)</i> Advisory Opinion OC-23/17 (15 November 2017)	433
<i>Velásquez Rodríguez v Honduras</i> Ser C No 4 (29 July 1988)	46–47

UNITED NATIONS

Committee on the Elimination of Racial Discrimination

<i>L.R. v Slovakia</i> Communication No 31/2003 (7 March 2005)	199
--	-----

International Court of Justice

<i>Application of the International Convention for the Suppression of the Financing of Terrorism and of the International Convention on the Elimination of All Forms of Racial Discrimination (Ukraine v Russian Federation)</i> (Provisional Measures Order of 19 April 2017)	199
<i>Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Qatar v United Arab Emirates)</i> (4 February 2021)	199

Human Rights Committee

<i>A v Australia</i> Communication No 560/93 (3 April 1997)	46–47
<i>Fijalkowska v Poland</i> Communication No 1061/02 (26 July 2005)	46–47

Table of Domestic Cases

AUSTRALIA

- Commissioner Initiated Investigation into Clearview AI, Inc (Privacy)* 54 AICmr
(Office of Australian Information Commissioner, 14 October 2021) 337

CANADA

- Nelson v Goodberry Restaurant Group Ltd dba Buono Osteria et al* [2021]
BCHRT 137 231
- Re Investigation into the Use of Facial Recognition Technology by the Insurance Corporation of British Columbia* BCIPCD No 5 (Office of the Information & Privacy Commissioner for British Columbia, 2012) 335–36

CHINA

- Supreme People's Court Regarding the Trial Use of Face Recognition Technology to Process Personal Information Provisions on Several Issues Concerning the Application of Law in Related Civil Cases* (Supreme People's Court, 28 July 2021) 337

FRANCE

- Friends of the Earth et al v Total Energy*, Nanterre High Court (2021) 522–23
- La Quadrature Du Net et autres*, Tribunal Administratif de Marseille, No 1901249 (27 February 2020) 141–42

INDIA

- Justice KS Puttaswamy (Retd) v Union of India* (2017) 10 SCC 1 152
- People's Union for Civil Liberties v Union of India & Others* (1997) 1 SCC 301 347

IRELAND

- Atherton v Director of Public Prosecutions* [2005] IEHC 429 185

ITALY

- Italian Constitutional Court decision no 149 (16 May 2008) 185
- Italian Constitutional Court decision no 135 (21 May 2012) 179
- Tribunal of Bologna Order no 2949/2019 (31 December 2020)* 378

MAURITIUS

- Hurnam v Paratian*, Privy Council on Appeal from the Court of Civil Appeal of Mauritius [1998] 3 LRC 36 296

NETHERLANDS

<i>NJCM cs v De Staat der Nederlanden (SyRI)</i> C-09-550982-HA ZA 18-388 (District Court of the Hague, 2020)	272, 355–56
Supreme Court App no 02632/02/04 [Cameratoezicht] (20 April 2004)	183

NEW ZEALAND

<i>Hosking v Runting Ltd</i> [2004] CA 101-03	185, 186, 187
---	---------------

NIGERIA

<i>Ntukidem v Oko</i> App no SC.30/1989 (Supreme Court of Nigeria, 12 February 1993)	296–97
---	--------

SWEDEN

<i>Skellefteå Municipality, Secondary Education Board Supervision Pursuant to the General Data Protection Regulation (EU) 2016/679—Facial Recognition Used to Monitor the Attendance of Students, Swedish Data Protection Authority, No DI-2019-2221</i> (20 August 2019)	141–42
---	--------

UNITED KINGDOM

<i>Campbell v MGN Ltd</i> [2004] UKHL 22	185
<i>Creation Records Ltd v News Group Newspapers Ltd</i> [1997] EWHC Ch370	186
<i>Douglas v Hello Limited</i> [2005] EWCA Civ 595	187
<i>R (Bridges) v Chief Constable of South Wales Police</i> [2019] EWHC 2341 (Admin)	201
<i>R (on the application of Bridges) v South Wales Police</i> [2020] EWCA Civ 1058	200–2
<i>Rylands v Fletcher</i> [1868] UKHL 1	496
<i>SM and Ihsan Qadir v Secretary of State for the Home Department</i> IA/31380/2014 (Upper Tribunal, Immigration and Asylum Chamber, 21 April 2016)	338
<i>United Biscuits</i> [1991] LON/91/0160	496–97

UNITED STATES

<i>California v Ciraolo</i> 476 US 207 (1986)	185
<i>Commonwealth v Robinson</i> , No CC201307777 (Pa Ct C P Allegheny City, 4 February 2016)	484–85
<i>Connecticut Fair Housing Center et al v CoreLogic Rental Property Solutions</i> 3:18-CV-705 (District Court of Connecticut, 2020)	360
<i>Dinerstein v Google</i> 484 F.Supp 3d 561 (United States District Court, ND Illinois, Eastern Division, 2020)	397–98
<i>Florida v Riley</i> 488 US 445 (1989)	185
<i>Goidel v Aetna Inc</i> Case No 1:21-cv-07619 (United States District Court Southern District Of New York, 2021)	227–28, 232–33
<i>Griswold v Connecticut</i> 381 US 479 (1965)	122
<i>Haff Poultry Inc et al v Tyson Foods Inc</i> No 6:17-CV-00033-RJS (US District Court for the Eastern District of Oklahoma, 2017)	351
<i>Katz v United States</i> 389 US 347 (1967)	122
<i>Kyllo v United States</i> 533 US 27 (2001)	122
<i>Loomis v Wisconsin</i> 881 N.W.2d 749 (Wis 2016)	58, 113
<i>Marsh v Alabama</i> 326 US 501 (1946)	87

<i>McCleskey v Kemp</i> 481 US 279 (1987)	287–88
<i>NetChoice v Paxton</i> 596 US __ (2022)	87
<i>Penny Quiteros, Plaintiff, v INNOGAMES et al, Defendants</i> No C19-1402RSM (WD Wash, 30 March 2022).....	415–16
<i>Prager University v Google LLC</i> 951 F.3d 9912 (2020)	87
<i>State v Loomis</i> 881 NW 2d 749, 755 (Wis, 2016).....	271–72, 278–79, 484–85

ZIMBABWE

Holland and Others v Minister of the Public Service, Labour and Social Welfare

App no ZLR 186 (S) (Supreme Court of Zimbabwe).....	296–97
---	--------

Abbreviations

A1P1	Article 1 of ECHR Protocol 1
AA	Artificial Agent
ACHPR	African Charter on Human and Peoples' Rights
ACHR	American Convention on Human Rights
ACommHPR	African Commission on Human and Peoples' Rights
ACRWC	African Charter on the Rights and Welfare of the Child
ADM	Automated Decision-Making
AFR	Automated Facial Recognition
AGI	Artificial General Intelligence
AGSI	Artificial General Super Intelligence
AI	Artificial Intelligence
AIA	Algorithmic Impact Assessment
AI Act	European Commission 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD)
API	Application Programming Interface
APPI	Act on the Protection of Personal Information (No 57 of 2003, as amended 2020) (Japan)
ATM	Automated Teller Machine
AV	Autonomous Vehicle
AWAVA	Australian Women Against Violence Alliance
AWS	Autonomous Weapon System
BAMF	Federal Office for Migration and Refugees (Germany)
BHR	Business and Human Rights
BIA	Business Impact Analysis
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CAS	Criminaliteits Anticipatie Systeem
CCPA	California Consumer Privacy Act (2018)
CCTV	Closed-Circuit Television
CDC	Center for Disease Control and Prevention
CEDAW	Convention on the Elimination of All Forms of Discrimination Against Women
CERD	Committee on the Elimination of Racial Discrimination
CESCR	Committee on Economic, Social and Cultural Rights
CFREU	Charter of Fundamental Rights of the European Union
CGI	Computer-Generated Imagery

XXX ABBREVIATIONS

CoE	Council of Europe
COHRE	Centre on Housing Rights and Evictions
COI	Country of Origin
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CRC	Convention on the Rights of the Child
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRPA	California Privacy Rights Act (2020)
CRPD	United Nations Convention on the Rights of Persons with Disabilities
CSES	Centre for Strategy and Evaluation Services
DICFP	Data-Informed Community-Focused Policing
DL	Deep Learning
DLT	Distributed Ledger Technology
DMA	Digital Markets Act
DPA	Data Protection Authority
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
DSA	Digital Services Act
DSS	Decision Support System
EA 2010	Equality Act 2010
ECHR	European Convention on Human Rights and Fundamental Freedoms
ECOWAS	Economic Community of West African States
ECtHR	European Court of Human Rights
ECommHR	European Commission of Human Rights
EDPS	European Data Protection Supervisor
EHDS	European Health Data Space
EHR	Electronic Health Records
EPC	European Patent Convention
EPO	European Patent Office
ESCAP	Economic and Social Commission for Asia and the Pacific
EU	European Union
eu-LISA	European Union Agency for the Operational Management of Large-Scale IT Systems
FAO	Food and Agriculture Organization
FCA	Financial Conduct Authority
FDA	US Food and Drug Administration
FEANTSA	European Federation of National Organisations Working with the Homeless
FGM	Female Genital Mutilation
FIPPs	Fair Information Practice Principles
FRA	Fundamental Rights Agency (EU)
FRT	Facial Recognition Technology
FSC	Food Supply Chain
GCHQ	Government Communications Headquarters (UK)
GDPR	General Data Protection Regulation

GIS	Geographic Information System
GLAAD	Gay and Lesbian Alliance Against Defamation
GPS	Global Positioning System
GPSD	General Product Safety Directive (EU)
HART	Harm Assessment Risk Tool
HFEA	Human Fertilisation and Embryology Authority (UK)
HIPPA	Health Insurance Portability and Accountability Act (US)
HRDD	Human Rights Due Diligence
HRIA	Human Rights Impact Assessment
HRIA-AI	Human Rights Impact Assessment in the context of Artificial Intelligence
HRW	Human Rights Watch
IACtHR	Inter-American Commission on Human Rights
IACtHR	Inter-American Court of Human Rights
ICCPR	International Covenant on Civil and Political Rights
ICE	United States Immigration and Customs Enforcement Agency
ICERD	International Convention on the Elimination of All Forms of Racial Discrimination
ICESCR	International Covenant on Economic, Social and Cultural Rights
ICJ	International Court of Justice
ICT	Information and Communication Technology
IEA	International Energy Agency
IFC	International Finance Corporation
ILO	International Labour Organization
IMCO	Internal Market and Consumer Protection
IoT	Internet of Things
IP	Intellectual Property
IPR	Intellectual Property Rights
IRCC	Immigration, Refugees and Citizenship Canada
ISO	International Organization for Standardization
IUI	Intrauterine Insemination
LAPD	Los Angeles Police Department
LAWS	Lethal Autonomous Weapon System
LGBTQI+	Lesbian, gay, bisexual, transgender, queer, intersex + other sexual orientation and gender identities
LMIC	Low- and Middle-Income Countries
MIT	Massachusetts Institute of Technology
ML	Machine Learning
NGO	Non-Governmental Organisation
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NRM	New Religious Movement
NSA	National Security Agency
NZ	New Zealand
OAS	Organization of American States

xxxii ABBREVIATIONS

OCHA	United Nations Office for the Coordination of Humanitarian Affairs
ODR	Online Dispute Resolution
OECD	Organisation for Economic Co-operation and Development
OHCHR	Office of the United Nations High Commissioner for Human Rights
OPSS	Office for Product Safety and Standards
PEPUDA	Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000
PIA	Privacy Impact Assessment
PIPEDA	Personal Information Protection and Electronic Documents Act 2000 (Canada)
PLD	Product Liability Directive
POST	Public Oversight of Surveillance Technology Act
Protocol of San Salvador	Additional Protocol to the American Convention on Human Rights in the Area of Economic, Social, and Cultural Rights
PWD	Persons with Disabilities
QR	Quick Response
Refugee Convention	Convention Relating to the Status of Refugees
RFID	Radio-Frequency Identification
RSD	Refugee Status Determination
SAMS	Social Assistance Management System
SDG	Sustainable Development Goal
SMM	Social Media Monitoring
SOGI	Sexual Orientation and Gender Identity
SWP	South Wales Police
SyRI	System Risk Indication
TFEU	Treaty on the Functioning of the European Union
TUC	Trades Union Congress
UDHR	Universal Declaration of Human Rights
UK	United Kingdom
UN	United Nations
UN Charter	United Nations Charter of 1945
UNHCR	United Nations High Commissioner for Refugees
UNCTAD	United Nations Conference on Trade and Development
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNGA	United Nations General Assembly
UNGPs	United Nations Guiding Principles on Business and Human Rights
UNHCR	United Nations High Commissioner for Refugees
UNHRC	United Nations Human Rights Committee

UNICEF	United Nations Children's Fund
UNSC	United Nations Security Council
URL	Uniform Resource Locator
US	United States
USBP	United States Border Patrol
VR	Virtual Reality
WEF	World Economic Forum
WFP	World Food Programme
WHO	World Health Organization
WOTF	Way of the Future
WP29	Article 29 Working Party
YORST	Youth Offending Risk Screening Tool

About the Contributors

Joe Atkinson is Lecturer in Employment Law at the University of Southampton

Matilde Ceron is Max Weber Fellow at the Robert Schuman Center of European University Institute

Ishita Chakrabarty is MA Candidate in International Law at the Geneva Graduate Institute

Ignacio Cofone is Canada Research Chair in AI Law & Data Governance at McGill University

Philippa Collins is Senior Lecturer in Law at University of Bristol

Caroline Compton is Senior Lecturer in Law at Western Sydney University

Giovanni De Gregorio is PLMJ Chair in Law and Technology at Católica Global School of Law

Sarah de Heer is PhD Researcher at Lund University

Paul De Hert is Professor of Privacy & Technology, Human Rights, and Criminal Law at Vrije Universiteit Brussel and Associated Professor at Tilburg University

Jan Drchal is Assistant Professor of Machine Learning at the Artificial Intelligence Center at Czech Technical University in Prague

Pietro Dunn is PhD Researcher at University of Bologna and University of Luxembourg

Isabel Ebert is Senior Research Fellow at University of St Gallen and Adviser to the UN Human Rights B-Tech Project

Silvia Favalli is Research Fellow in International Law at University of Milan

Florian Gamper is Research Associate at Singapore Management University

Valentina Golunova is PhD Researcher at Maastricht University

Elizaveta Gromova is Deputy Director for International Cooperation and Associate Professor at National Research South Ural State University

David Gunkel is Professor of Communication Technology at Northern Illinois University

Klaus Heine is Professor of Law and Economics at Erasmus University Rotterdam

Jessie Hohmann is Associate Professor at University of Technology Sydney

Lisa Hsin is Helsby-Kroll Postdoctoral Research Fellow in Business and Human Rights at the Bonavero Institute of Human Rights and Junior Research Fellow at Corpus Christi College, University of Oxford

Malcolm Katrak is Assistant Professor at Jindal Global Law School and a York Graduate Fellow at Osgoode Hall Law School

Louis Koen is Assistant Lecturer in Law at University of Johannesburg

Luboš Král is Deputy Director of the Artificial Intelligence Center at Czech Technical University

Migle Laukyte is Tenure Track Professor in Cyberlaw and Cyber Rights at Pompeu Fabra University

Shu Li is Postdoctoral Researcher at the Legal Tech Lab at University of Helsinki

Fabian Lütz is PhD Researcher at Université de Lausanne

Lotta Majewski is Licensed Legal Counsel at Asianajotoimisto Asianaiset Oy Helsinki

Natalia Menéndez González is PhD Researcher at European University Institute

Helga Molbæk-Stensig is PhD Researcher at European University Institute

Kgomotso Mufamadi is Lecturer in Law at University of Johannesburg

Adekemi Omotubora is Senior Lecturer at University of Lagos

Alessandro Ortalda is PhD Researcher at Vrije Universiteit Brussel

Marília Papaléo Gagliardi is Digital Rights Advisor at Article 19

Andrea Pin is Associate Professor of Comparative Public Law at University of Padua

Kostina Prifti is PhD Researcher at Erasmus University Rotterdam

Alexandre Quemy is PhD Researcher at Poznan University of Technology

Alberto Quintavalla is Assistant Professor of Innovation of Public Law at Erasmus University Rotterdam

Sofia Ranchordás is Professor of Administrative Law at University of Tilburg and Professor of Law and Innovation at LUISS Guido Carli

Raimy Reyes is a Human Rights Lawyer specialised in vulnerable groups in the context of human mobility

Enrique Santamaría Echeverría is Postdoctoral Researcher in Law at Erasmus University Rotterdam

Béatrice Schütte is Postdoctoral Researcher at the Legal Tech Lab at University of Helsinki and University of Lapland

Masuma Shahid is Lecturer in EU Law and PhD Candidate at Erasmus University Rotterdam

Martina Šmuclerová is Senior Lecturer in Public International Law at Sciences Po Paris and Senior Research Fellow at Ambis University in Prague

Dhruv Somayajula is Research Fellow at the Centre for Applied Law and Technology Research at Vidhi Centre for Legal Policy

Evert Stamhuis is Professor of Law & Innovation at Erasmus Universiteit Rotterdam

Jeroen Temperman is Professor of International Law at Erasmus University Rotterdam

Letizia Tomada is Researcher at the Centre for Information and Innovation Law at University of Copenhagen

Bart van der Sloot is Associate Professor at Tilburg University

Margaret Warthon is PhD Researcher at University of Groningen

Raphaële Xenidis is Assistant Professor in EU Law at University of SciencesPo

Antonella Zarra is PhD Researcher in Law and Economics at Hamburg University

Alessia Zornetta is SJD Candidate at University of California Los Angeles

PART I

AI-BASED HUMAN RIGHTS VIOLATIONS: LEGAL AND TECHNICAL BACKGROUND

1

Introduction

Alberto Quintavalla and Jeroen Temperman

1 Artificial Intelligence and Human Rights: Parallel Universes

Both early artificial intelligence (AI) milestones and the modern human rights codification process have their origins in the 1940s.

In the 1940s, important early AI foundations saw the light of day, including Warren Sturgis McCulloch and Walter Pitts' *A Logical Calculus of the Ideas Immanent in Nervous Activity* (1943), laying down the foundations for Turing-complete artificial neurons. In 1950, Alan Turing ventured—in a paper entitled 'Computing Machinery and Intelligence' that was published in *Mind*—into the famous question, 'Can machines think?' The ensuing Turing Test serves to establish precisely that, whether or not a non-human entity—a computer, machine, or robot—'can think'.

On 10 December 1948, the General Assembly of the newly founded United Nations (UNGA)—created as per the 1945 UN Charter and aiming to prevent the atrocities the League of Nations helplessly failed to avert—adopted the Universal Declaration of Human Rights (UDHR), positing that the 'recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world' (Preamble). The UDHR, consisting of a combination of civil and political rights on the one hand, and social, economic, and cultural rights on the other, serves as the milestone contemporary human rights instrument up until today, also influencing numerous subsequent international and national bills of rights.

Their overlapping history notwithstanding, the two phenomena—AI and human rights—led fairly separate existences for their first fifty or sixty years or so. It is only in the last decade that their paths have converged, that the two forces meet, that they support each other, or, as may happen as well, that they conflict, causing small or major clashes.

2 The Rapprochement of Artificial Intelligence and Human Rights: Friends or Foes?

Artificial intelligence and human rights are currently interacting within one and the same world and their inevitable dynamics no longer go unnoticed, though those dynamics simultaneously pose tremendous—and tremendously pertinent—legal, ethical, technological, and societal questions.

The term ‘artificial intelligence’ is a broad concept encompassing multiple applications that do things requiring intelligence. All these applications are likely to have a significant impact on human rights. On the one hand, they may contribute to the advancement of human rights. For instance, the use of machine learning (ML) in healthcare could improve precision medicine and eventually provide better care to patients. On the other hand, they can pose an obvious risk to human rights. One needs only to think of the many instances where biased algorithms discriminate against ethnic minorities and women.

This double-edged sword understanding of AI has swiftly become the dominant paradigm, also within political and legislative arenas. In the past decade, international organisations—including their ranking human rights officers—have been tumbling over each other to flag the issue of AI, typically arriving at the conclusion that, from a human rights perspective, AI presents both benefits and risks. However, how precisely these impacts pan out under the different fundamental rights; and, more specifically, how these dynamics play out *amongst* the various rights, and various categories of rights, has been somewhat of a black box—that is the void the present book aims to step into.

3 Scope and Objectives

Within international fora, pioneering benchmarking has gradually commenced in the form of guidelines and recommendations at both international and regional levels (see eg the UN guidelines, European Union (EU) Act and—guidelines, Council of Europe (CoE) guidelines, and numerous other recommendations on AI that amply feature and are engaged with throughout this volume). While those do—as will amply be illustrated in the pages that follow—provide some level of detail, typically the significance of the relationship between AI and human rights, as analysed and understood in those guidelines, remains limited to a number of areas of particular concern, notably: data protection, privacy, equality or non-discrimination, procedural fairness, and questions of accountability. While those angles engage a number of human rights rather directly, many human rights standards beyond such usual suspects as private life, justice rights, and equal treatment rights remain obscure in terms of their dynamics with AI.

This is not all that surprising. The explanation lies arguably in a mixture of prioritisation and the quest for clarity. That is, looking at the most likely impact of AI on human rights, data protection, privacy, and fair treatment obviously is not a bad start—those are fundamental rights issues clearly on the frontline of AI-driven interferences. The direct impact of AI on certain other human rights may, on its face, be more difficult to chart, predict, or even to comprehend altogether.

This volume aims to warn against complacency and to show that AI impacts, or will imminently impact, human rights across the board. Here, the term ‘impact’ is used neutrally, for in addition to adversaries, human rights and AI at times and under the right regulatory conditions operate in a friendly manner, with the two phenomena mutually reinforcing each other’s outputs. Indeed, technological innovation and human rights promotion can, under the right circumstances, go hand in hand.

This volume also ventures into the dynamics, whether beneficial or adverse, between AI and virtually all types of human rights, including vast samples from the category of first generation rights, civil and political rights, as well as second generation rights—economic, social, and cultural rights—and also third generation rights—collective solidarity rights. Naturally, given the volume’s focus, the newly emerging fourth generation of rights—not fully crystallised yet but which may be referred to as digital access, digital self-determination, and digital security, including *habeas data*—will also be amply engaged throughout the thematic human rights studies and in the final reflections.

Hence, this volume’s mission is to chart AI and human rights dynamics comprehensively. For many substantive human rights, this volume’s engagement from an AI viewpoint is ground-breaking: these dynamics have either not yet been reported in the literature, or the latter has scraped the surface of reportable risks or benefits. In any event, the chapters that comprise this volume seek to systematically, comprehensively, and inclusively chart and predict AI’s dynamics with the substantive right—be it first, second, or third generation—at hand, keeping an eye out for both positive relations (AI as friend to the right at stake), as well as adverse ones (AI as detrimental to the right at stake). In respect of the latter, the chapters also seek to pinpoint—as far as possible—where precisely in the AI life cycle damage or risks to human rights are detectable or likely, ranging from design phases to AI deployment phases.

That said, the ‘usual suspects’ are usual suspects for a reason: here, AI’s impact is—for the time being—the most visible and the most predictable. Accordingly, this collection certainly does include introductions to such themes of AI and privacy, AI and fair procedure, and AI and non-discrimination. At the same time, we aim to chart the bits of uncharted ground within these otherwise charted territories, as it were. For instance, in addition to racial profiling and racist usage of AI, how about impact on LGBTQI+ rights? And when it comes to privacy matters, are AI-based encroachments universal, or may we detect differing impacts, for instance,

between the Global South and Global North? There are enough pertinent matters to investigate also within these areas, what with AI developments being relatively novel and, at the same time, moving so swiftly, and also considering AI's black box problems as amply reported and illustrated throughout this volume.

The authors of the 'individual human rights' chapters have all been expressly commissioned to take a specific substantive right as concrete point of departure and to build their risk and benefit assessment upon the dynamics between AI and the right in question. However, it would be legally and analytically wrong and unwise—also from the perspective of overcoming any challenges flagged—to present such dynamics studies in total isolation. The interdependence and indivisibility of human rights imply that effects on the one right rub off on the other. This may cause negative spirals of rights enjoyment, as widely reported in these chapters. But, at the same time, indivisibility may also be the key to remedying at least some risks posed to fundamental rights. As a result, this holistic approach to contemporary human rights has left imprints on many of these chapters, which typically include a section on 'related rights' impacts, while the numerous internal cross-references serve to complete the highly dynamic picture.

After zooming in, zooming out is necessary, not only to grasp again such questions of interdependence better, but also to take stock of where does AI bring us, as far as our human rights enjoyment is concerned? Binary and absolute (AI bad, AI good) conclusions are unlikely to do justice to the richness and nuances presented by the substantive case studies. Zooming out, hence, also means approaching the same subject matter—human rights and AI—from multiple perspectives and disciplines. Consequently, we have selected a number of legal theory, moral philosophy, and epistemology accounts that foster stocktaking. Likewise, we included some contributions offering a more analytical discussion on general AI-related issues such as human rights impact assessment (HRIA) or the crucial role of the business sector in the context of AI.

4 Outline

We have organised this volume by subdividing the materials into ten parts, starting with this first part that lays the legal-technical groundwork for critically analysing the dynamics between AI and human rights, moving on to the volume's core parts containing thematic human rights studies, ranging from AI dynamics with first, second, and third generation rights (Parts II–VIII), and ending the volume with a part containing multidisciplinary and bird's-eye reflections on AI and human rights (Part IX) and, finally, a concluding part (Part X). Let us introduce the volume's specific parts in more detail:

4.1 Part I

In what follows in this first part of the volume the scene is further set by providing what is essentially a genesis of AI-based risks to human rights. In what ways can AI adversely impact human rights and, specifically, how do those threats emerge within the process of developing AI? Especially for all human rights stakeholders unfamiliar with the technical ins and outs of AI, a team of AI experts who are also well versed in human rights terminology—Martina Šmuclerová, Luboš Král, and Jan Drchal—unpack the issue of ‘AI and human rights’ from a technical perspective and show how human rights issues materialise within the AI life cycle. We learn there is no singular threat stemming from AI, rather the root causes for AI-based human rights violations may derive from any and all of the stages of AI. The various phases—be it input, transfer, or deployment—all come with their unique threats, sometimes in the form of malicious abuse, but oftentimes violations and risks are due to human errors somewhere in the design or application phases. Chapter 2 on the AI life cycle, hence, has a dual function: first of all, it serves to contextualise, qualify, and gauge the human rights concerns in the volume’s subsequent thematic case studies (are we dealing with data issues? input issues? transfer issues? or rather deployment issues? and so on). Second, the chapter helps to contemplate, in a general way, potential forms of human rights violation prevention and redress, namely—in addition to structural and systemic safeguards like human rights risk assessments and transparency—in the form of tailor-made approaches that are cognisant of the AI life cycle.

4.2 Part II

Part II analyses the dynamics between AI and assorted first generation—civil and political—rights. In chapter 3, Valentina Golunova shows how various recently deployed AI systems, notably in the area of policing, criminal justice, and surveillance, affect the right to liberty and security. In chapter 4, Jeroen Temperman deploys the tripartite respect, protect, and fulfil duty discourse to illustrate that AI and religious freedom impact each other. While, at times, these effects are beneficial, pertinent questions to relevant AI-designs and use cases are raised while positive obligations on the part of the state come to the fore to prevent or remedy encroachments on the right to freedom of religion or belief.

While automated technologies have the potential of fostering access to information, AI-based faulty or biased content moderation and content prioritisation undermines free speech, information rights, and also puts pressure on other civil and political rights. Accordingly, in chapter 5, Giovanni De Gregorio and Pietro Dunn show how freedom of expression and information in the digital age reconfigures protection needs, urging a calibration of the state’s positive duties to foster

compliance with this cornerstone democratic right, in addition, that is, to the need to place enhanced emphasis on duties vis-à-vis and directly borne by private media and information platforms. Essentially, the private governance of free speech and information calls for re-politicisation, with the language of such democratic discourse being informed by human rights standards, notably free speech, access to information, as well as demands from pluralism. AI's adverse impact on another fundamental right also comes in the form of chilling effects. In chapter 6, Margaret Warthon shows the use of biometrics in the public square, especially their emerging role in 'monitoring' protests. Far from being the infallible, neutral, and unbiased AI systems that developers or users claim them to be, their deployment in many contexts fails to satisfy the necessity and proportionality principles and causes breaches of the freedom of assembly.

In chapter 7, Letizia Tomada and Raphaële Xenidis venture into the dynamics between AI and property rights, especially with a timely focus on intellectual property, thus arguably navigating across the various generations of rights, from first (property per se), to second (property as economic freedom), to third (common heritage of mankind-style assets), and even fourth generation—digital era-related—rights. First, they address the issue of whether AI itself can be subject to property rights protection, before moving on to illustrating the impact (property rights-protected) AI may have on human rights, highlighting the freedom of expression and information for being particularly at risk, while also victims of algorithmic discrimination are—doubly—adversely affected as is visible in litigation cases where the latter's burden of proof is prohibitively high. A combination of a reconsideration of exceptions within the intellectual property law framework and—to the extent that property rights discourse is applicable—a better deployment of the limitation techniques and proportionality tests within the human rights framework is proposed to redress the flagged risks.

4.3 Part III

Notwithstanding the stated objectives to cover as much novel ground as possible, some dynamics and correlations are hotly debated for a reason—they currently emerge as the direst or most challenging or at least the most visible ones. Our mission in those regards is twofold, namely (i) to present solid introductions to these matters for the lay person, and (ii) to aim to enrich the discussion by approaching the challenge at stake from a multitude of perspectives. Accordingly, Part III deals with privacy rights. In chapter 8, Alessia Zornetta and Ignacio Cofone flesh out the underlying principles, the legal rationales behind privacy and data protection in our digital age, specifically applying these same principles to the development, application, or control of privacy-related and potentially interfering forms of AI. Part III shows not only that numerous AI technologies—including facial recognition

technologies (FRT), among other surveillance measures—affect privacy, but also how privacy in our digital age is a cornerstone right.

The latter means that privacy interferences negatively rub off on numerous other civil and political rights as well as on second generation rights. In chapter 9, Natalia Menéndez González further illustrates this corollary function of privacy rights and data protection in our digital age, using FRT as a case study. Such an umbrella function hinges strongly on checks and balances and other rule of law safeguards. Negatively, the inverse corollary function holds true *a fortiori*. In chapter 10, Malcolm Katrak and Ishita Chakrabarty demonstrate significant negative correlations between privacy breaches and violations of civil and political rights within important parts of the Global South. Privacy encroachments are reported as stepping stones towards stifling political participation and dissent. Such digital authoritarianism can only be properly addressed if we redefine privacy as a comprehensive civil-political notion.

In chapter 11, Bart van der Sloot identifies the two dominant ideal-typical strategies towards privacy and data protection: individualised control versus government regulation. Dismissing both as inadequate in our age of AI, chiefly on account of the impossibility of consent and insurmountable digital power imbalances, he suggests that we need to let go of the possibility of ‘controlling’ data; instead we must intensify our focus on the initial knowledge production process. In chapter 12, Andrea Pin shifts the entire privacy paradigm yet further, advocating anonymity rights and—importantly—reconceptualises what makes up ‘the public sphere’ in the digital age.

4.4 Part IV

Part IV focuses on questions of equality and non-discrimination as raised by AI and its design or current usages. In chapter 13, Louis Koen and Kgomoitso Mufamadi dispel the illusion that AI would be ‘colour blind’. AI, as currently deployed, serves to entrench racial discrimination and economic and other disparities. A combination of the state’s positive human rights obligations and the principle of business due diligence are highlighted with a view towards redressing ongoing forms of AI-based racial discrimination.

From representation issues within tech companies, to biased training data sets, to the perpetuation of racist biases in the deployment phase, AI serves to entrench marginalisation of historically underprivileged and vulnerable groups—a notion observed and reported and illustrated, it may be added here, throughout the volume. Accordingly, the design and application of AI technologies—if unguided, unrestricted, or unregulated—also affects women’s rights. In chapter 14, Fabian Lütz shows how algorithms entrench and exacerbate gender-based discrimination, often due to biases within the training datasets, while Marília Papaléo Gagliardi’s

chapter 16 on deepfake violence against women deconstructs another specific contemporary and particularly misogynist (ab)use of AI.

AI shows certain potential for the promotion of LGBTQI+ rights, notably in such areas as health and security. Otherwise, similar adverse dynamics as narrated in the women's rights and racial equality chapters hold true: biased algorithms further entrench and perpetuate marginalisation and discrimination. Using a variety of case studies in chapter 15, Masuma Shahid aims at overcoming such AI systemic oversights and biased data sets by queering AI, thus ensuring that AI is SOGI (sexual orientation and gender identity) inclusive from the drawing board through to the deployment of automated technologies.

In chapter 17, Antonella Zarra, Silvia Favalli, and Matilde Ceron discuss digital inclusion and non-discrimination of persons with a disability. Using the United Nations Convention on the Rights of Persons with Disabilities (CRPD) as key yardstick, selected use cases show how AI fosters and significantly hampers inclusive equality. Entrenching existing bias through digital exacerbation, once again, is a striking blemish on AI as currently designed, trained, tested, and deployed, forcing a re-conceptualisation of AI systems and their usage based on a disability human rights-based approach to AI.

4.5 Part V

Part V builds on the discrimination narrative by discussing, in a comprehensive manner, issues of fair procedure, including fair legal procedure, legal remedies, and miscellaneous procedures outside of the courtroom. While these questions are inextricably linked to the previous equal treatment discussion—the design or uses of AI causing forms of discrimination—the commonality of these particular case studies lies in their concerns as to the human rights principle of fair, transparent, reasonable, and equitable treatment.

In chapter 18, Helga Molbæk-Stensig and Alexandre Quemy present an overview of the right to a fair trial per se, being the most violated fundamental human right and at the same time, potentially key to redressing rights interferences, including AI-based violations. Charting the vacuum where AI may step in from a perspective of human fallibility (human judges are not perfect), they also comprehensively outline the risks that AI poses to fair trial rights if unregulated.

In chapter 19, Migle Laukyte unveils the dark side of far-reaching data analytics, both inside the court room as well as within healthcare. In addition to privacy and personal data abuses, discriminatory practices and interferences with personal autonomy are highlighted through the chapter's case studies.

AI-driven decisions that negatively affect individuals may, designedly or inadvertently (on account of their code's sheer complexity), be difficult to grasp both by the affected persons themselves and by the judiciary. In chapter 20, Sarah de Heer

shows that AI-based automated decision-making (ADM) risks, unless adequately regulated, upsetting one of the cornerstones of human rights implementation: effective remedies.

4.6 Part VI

Article 14 of the UDHR provides that: ‘Everyone has the right to seek and to enjoy in other countries asylum from persecution’. Part VI, focusing on the right to asylum, aims to unpack the dynamics between emerging AI technologies and this fundamental right. While Dhruv Somayajula’s chapter 22 focuses on the screening of refugees specifically, showing how AI has been adopted within the area of ‘migration management’ for reasons of efficiency and warning about the risks stemming from AI-based analytics, Raimy Reyes in chapter 21 comprehensively analyses the entire cycle of forced migration and shows the benefits but especially the risks posed by AI to the right to asylum. She reminds us that these ‘technologies have been traditionally developed by the private sector for surveillance control purposes and have historical antecedents in colonial technologies of racialised governance. Thus, these AI technologies are not neutral, and their design and use typically reinforce dominant social, political, and economic trends’.

4.7 Part VII

In Part VII, we move on to the second generation of fundamental rights: economic, social, and cultural rights. In chapter 23, Adekemi Omotubora engages with the right to food and illustrates the positive impact AI may have on equal access to food through smart farming, anchoring agricultural precision and optimisation AI-technologies within the state’s duty to respect, protect, and provide.

In chapter 24, Caroline Compton and Jessie Hohmann explain the implications of AI for the right to housing, showing that while most novel technologies may appear neutral, their deployment occurs in a context of historical inequalities and vulnerability, thus risking to exacerbate marginalisation and disparities in the housing area. AI certainly comes with a degree of potential for promoting housing rights—including disaster risk reduction, improving housing rights for elderly and disabled persons, and enhanced sanitation—whilst the general human rights framework ought to be utilised to tackle the flagged negative dynamics.

Significantly affecting first generation rights, too, such as privacy, free speech, freedom of association, and non-discrimination—all of which are also important ‘rights at work’—Joe Atkinson and Philippa Collins show how AI, and specifically ‘algorithmic management’ impact the right to work specifically, including the right to decent working conditions in chapter 25. They argue ex post redress attempts

fail to take workers' rights threatened by AI seriously and formulate various *ex ante* measures to pre-empt encroachments, including through risk assessments and collective bargaining over the very use of these novel technologies within the labour context.

In chapter 26, Enrique Santamaría Echeverría charts the most significant AI-based developments in healthcare, including in such personal and collective health areas as diagnosis and disease identification, personalised treatment, mental health, digital phenotyping, public and population health, translational research, clinical care, and healthcare management. In so doing, he illustrates the potential for the promotion of the right to health as well as the risks and challenges—notably on account of AI's inaccuracy, unexplainably, and opacity resulting from privacy and data protection questions, as well as cybersecurity issues—to the same right and various other rights that are simultaneously implicated. A separate focus on policy and regulatory recommendations serves to steer away from these latter adverse impacts.

4.8 Part VIII

Part VIII is the final part that revolves around designated thematic human rights studies of AI impact and potential and is, more specifically, focused on third generation rights. This emerging and contemporary body of rights includes collective solidarity rights and, therefore, seeks to move beyond the classical human rights paradigm pointing to the individual as the rights holder or to one particular state as the duty bearer. These are rights we enjoy as global citizens, and that we enjoy *erga omnes* vis-à-vis the world community of states, the right to peace, sustainable development, and a socially just world order being striking examples. In the AI era, particularly engaged are consumer rights and environmental rights.

Accordingly, Shu Li, Béatrice Schütte, and Lotta Majewski illustrate in chapter 27 how AI can enable or destroy consumer rights. AI's potential is in this regard—unlike many other rights as the thematic case studies show—on paper, and in practice, already widely visible. AI as a force for good in the context of consumer protection is, for instance, linked to enhancing informed and protected internet usage, including online shopping, while benefits are also promising in the area of online dispute settlement where AI technologies foster consumers' fair procedure rights and access to remedies. Once again, the double-edged sword nature of AI is fully applicable. In the context of consumer protection rights, the authors narrate examples of AI-based consumer manipulation, product unsafety, and consumer discrimination. As the common denominator underlying those risks is formed by the fact that in the world of big data and algorithms a major knowledge schism has emerged between buyer and seller, most of their legal reforms suggestions seek to close that gap and overcome the disadvantageous position of consumers.

In chapter 28, Alberto Quintavalla considers Earth-friendly uses of AI including harmful substance spillage detection and energy optimisation among various other efficiency gains, as well as alarming—from an environmental perspective—designs and deployments of AI. This latter discussion, in addition to identifying various forms of application-based abuses, does not shy away from conceptualising AI for what it is when stripped down, environmentally speaking, to its bare core: a pollutant, for AI heavily relies on energy sources that are far from carbon neutral at present. Engaging with the emerging right to a healthy environment, chapter 28 considers redress and mitigation strategies, taking into account the complications posed by currently dominant state obligations versus weakly developed corporate obligations.

4.9 Part IX

After this mapping exercise that consists of zooming in on individual human rights and AI, it is time to zoom out in Part IX, which contains more reflexive pieces. While the authors of the individual human rights chapters were tasked to keep their antennae out for any dynamics among various rights and freedoms in the AI context, it is only by zooming out—and on the basis of a multitude of scientific disciplines—that we can take stock of AI and where it brings us in terms of human rights promotion and compliance. Breaking the ice in this regard, Kostina Prifti, Alberto Quintavalla, and Jeroen Temperman seek to digest the chief findings of this volume and at the same time engage in a forward-looking regulatory and policymaking exercise in chapter 29. They argue that in terms of common risks and benefits, stemming from the emerging ‘AI and human rights’ field, we need to distinguish between more structural and more functional risks—besides the modestly reported benefits—with each type of risk, upon close inspection, calling for its own tailored regulatory and/or technical response. Moreover, such responses engage action at different stakeholder levels, ranging from the international to the local.

Such bird’s-eye views, doctrinal and philosophical contributions, and interdisciplinary accounts also include investigations into such systemic and foundational questions as legal personality, the nature of decision-making, or the actual ‘smartness’ of so-called smart use cases and the concomitant role of public law; as well as the larger regulatory and resolution queries, including questions about business and human rights regulation, the need for human rights impact assessments, and human rights-conducive experimentation with AI.

In chapter 30, Klaus Heine addresses legal personality and AI. From an epistemological perspective, he argues that it is consciousness that distinguishes humans from AIs and as long as our common understanding is that AIs at best mimic our decision-making, this would stand in the way of legal personality, rights-holdership, and direct accountability. Heine engages the question what if

deep AI does become a reality? from a moral philosophy perspective, unveiling the paradox that humans may well have self-serving and pragmatic reasons to accrue legal personality beyond the human context. With AI as potential ‘top dog’, respect for human rights is after all best scripted into the conscience of the AI personality. A bit of a trade-off, but if there is going to be personality then preferably a human (rights) friendly personality. In chapter 31, David Gunkel, to the contrary, argues that the human rights discourse is an altogether faulty point of departure; rather, the question is if and to what extent robots or AIs possess the qualities or properties to be deemed ‘moral subjects’. If so, rights (natural, not human), but crucially also responsibilities, are engaged. Natural rights discourse, however, as Gunkel ambiguously concludes, contains the strongest arguments both in favour of and against robot rights.

Another chief undercurrent in the AI debate is the fundamental question what decisions or decision-making processes can be contracted out to machines—and which ones not. In chapter 32, Florian Gamper presents a case against AI involvement in normative areas, while his thematic case studies illustrate that this is precisely one of the arenas where AI is vastly emerging—and causing discrimination or other unjust treatment.

In chapter 33, Sofia Ranchordás warns that smart cities entrench marginalisation, bias, discrimination, and exclusion. Whereas public law could traditionally be geared towards redressing most types of abuses, forms of harm, and human rights violations; in the AI era, public law loses touch and lags behind the intricate nature of contemporary harms. Addressing this dissonance between smart urban solutions and the logics and modus operandi of public law, and the various forms of exclusion caused by smart urban planning, a reshaping of the protective qualities of administrative and other public law is required.

Further on the required regulatory framework that could face the risks stemming from AI, Isabel Ebert and Lisa Hsin comprehensively portray the state of the legal art as far as the business and human rights framework is concerned in chapter 34. First casting the net wide in terms of drawing the generally emerging, largely UN-steered business and human rights regime, they specifically consider private sector responsibilities in the age of AI, outlining that the previous ‘corporate irresponsibility gap’ is being closed and benchmarks for tech companies to prevent and redress human rights violations are mounting. They argue that legislation urging business human rights due diligence obligations, applicable to the entire AI life cycle from design to deployment, is the most promising route from a human rights promotion perspective.

Šmuclerová, Král, and Drchal intuited in chapter 2 that human rights impact assessments, covering all relevant phases of the AI life cycle, are a key to risk prevention and human rights promotion. We resume this discussion in this final part with Alessandro Ortalda and Paul De Hert fleshing out what such an assessment should look like in chapter 35. Finding inspiration in the data protection impact

assessments as mandated by European legislation, they underscore that a full-fledged five-step ‘Artificial Intelligence Human Rights Impact Assessment’ ought to go beyond data protection per se, urging sensitivity to the entire *acquis* of human rights.

In chapter 36, Elizaveta Gromova and Evert Stamhuis investigate the use and value of regulatory sandboxes in relation to AI and human rights. Moulding their usage to the precautionary principle and the positive and negative human rights obligations and other requirements flowing from international law, they seek to provide guidance on sandboxing practices that are legally resilient and human rights conducive. Concluding that no outright bans on regulatory sandboxes may be distilled from human rights law (HRL), they contend that under the right terms and conditions, real-life experimentation with AI-powered systems could serve to benefit all, from designers, regulators, to—most importantly—individuals.

4.10 Part X

Finally, Part X consists of the concluding reflections on the part of the editors.

2

AI Life Cycle and Human Rights

Risks and Remedies

*Martina Šmuclerová, Luboš Král, and Jan Drchal**

1 Introduction

The relationship between artificial intelligence (AI) and human rights is, for now, a double-edged sword. AI brings great benefits to all sectors of society and strengthens progress, social well-being, and economic competitiveness via automatisation of the respective human activities. At the same time, however, it poses risks to a variety of human rights and fundamental freedoms, be it due to the intrinsic technological processes, human input, or its abusive or malicious use in practice.

In order to assure the effective implementation of human rights norms in the AI domain, it is first necessary to identify the root causes of the human rights violations within the AI life cycle. These root causes can reside in all phases of the AI life cycle, starting with the incomplete input data, passing by a biased transfer learning, up to a malicious application. Likewise, the remedies can be diverse: problems might be solved via, for instance, a technical adjustment of the machine learning (ML) procedure, rules on the processing of data, or a more robust legal interference restricting or banning the development and use of certain AI technologies. A comprehensive approach linking international human rights and AI expertise is thus indispensable in order to provide a holistic and solution-oriented viewpoint.

This chapter follows up the contemporary discussion that has often focused on the impact of particular AI technologies on specific human rights¹ by providing an overall assessment of the AI life cycle based on an interdisciplinary approach. As such, it presents the systematisation of human rights violations throughout the whole AI life cycle, including the identification of the root causes and formulation

* This chapter presents the results and recommended remedies based on an interdisciplinary research done by international lawyers and AI experts in the framework of a grant research project 'AI and Human Rights: Risks, Opportunities and Regulation', supported by the Technological Agency of the Czech Republic (2021–23, Project No TL05000484).

¹ See eg Frederik Zuiderveen Borgesius, 'Strengthening Legal Protection Against Discrimination by Algorithms and Artificial Intelligence' (2020) 24 International Journal of Human Rights 1572; Kalliopi Terzidou, 'The Use of Artificial Intelligence in the Judiciary and Its Compliance with the Right to a Fair Trial' (2022) 31(3) Journal of Judicial Administration 154.

of a remedy mechanism. Although different human rights might be at risk in various fields of AI deployment, the underlying intrinsic processes and core technical elements of AI technology are common and similar factors and deficiencies could provoke human rights violations.

The chapter outlines various stages of AI systems under this human rights perspective. The ML-based system serves as a blueprint for the presented AI life cycle; other classical AI approaches such as planning, agent-based approaches, or game-theoretical methods typically involve only a subset of such stages. The AI life cycle encompasses steps and activities that lead to the design, development, and deployment of the AI product, verification of its functionality and its use as well as the solution of issues arising during its operation. The chapter is structured according to the phases of the AI life cycle: business understanding (section 2); data preparation (section 3); modelling (section 4); evaluation (section 5); and deployment (section 6).² The final section introduces certain requirements to enhance the human rights compliance that apply to the whole AI life cycle (section 7).

2 Business Understanding

Business understanding is the initial stage of the AI life cycle and in this phase information about the domain of deployment of the AI technology is collected. In other words, this stage aims at addressing the question: what does the business need? For example, as regards an AI technology used in healthcare for a disease diagnosis, it is necessary to collect all possible examples of the disease symptoms, its existing diagnoses, patients' data, data about the diagnostic process and the diagnosis delivery, as well as regulatory requirements. All actors of the AI life cycle—the producer, supplier, customer,³ and the regulator—may participate in this stage in order to perform the input analysis and determine the user requirements, the so-called 'user stories'.

The input analysis assesses the available data and current conditions of the AI system target domain. The user requirements define, for example, the use-cases (the services the AI system will provide), deployment, operation specifications (eg infrastructure or personnel needed to operate the system), and parameters of the system sustainability. The user requirements translate, *inter alia*, into technical

² These constituent stages of the AI life cycle correspond to the standardised methods such as the Cross-Industry Standard Process for Data Mining (CRISP-DM) or the Team Data Science Process. See eg Colin Shearer, 'The CRISP-DM Model: The New Blueprint for Data Mining' (2000) 5 *Journal of Data Warehousing* 13; or the Team Data Science Process, see eg Microsoft Azure, 'What is the Team Data Science Process?' <<https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview>>. In this article, the last stage—deployment—covers both deployment and operation (while some of the referred standardised methods, eg CRISP-DM, do not cover the operational stage).

³ The customer who acquires the AI system includes the operator who runs the AI system and the end user.

specifications. The specifications of the AI system contain the information about the application domain, the requested features of the AI technology, its architecture, technical design, and the deployment conditions.

It is in the user requirements and the initial analysis where a great portion of potential risks of human rights violations can be prevented. The limitations applicable throughout the whole AI life cycle with respect to human rights should be set in this initial stage. The supplier or the user can, for example, request that the AI system, designated to hire new employees, operates with no data indicating ethnicity or gender in order to maintain it unbiased and based purely on professional expertise. Similarly, if an autonomous vehicle (AV) is to be deployed also in a country with specific traffic signs (eg EU countries have different shapes and tones of colours for priority signs and warning signs, including typefaces in text), such a context must be determined in the user requirements so that its deployment would not put human lives at risk.

In other words, an important volume of human rights violations arises due to the insufficient determination of the user requirements and their translation into technical specifications, and thus an insufficient definition of the framework of the human rights limitations. Many of the following risks arising in subsequent stages of the AI life cycle could have been prevented in this initial stage (see section 7.1).

3 Data Preparation

The data preparation phase encompasses the collection of development data in raw form, data cleansing and transformation, the definition of attributes, training and testing data splits, labelling, and a full-fledged data shaping to be ready for the modelling phase.

Development data, meaning all data aimed to build the ML model, are collected in raw form and as such they enter the AI life cycle. For example, data for a predictive justice system may come from police, investigation, and court databases, while data for ad targeting might come from sources published on the Internet or even public census record databases covering demography statistics.

Data is cleaned and transformed. For example, erroneous and incomplete records are removed and missing pieces of data are supplemented where necessary. In that regard, care shall need to be taken to ensure that the information value carried by the data is not compromised. The first key moment of a human interference into the information value of the development data is the definition of attributes—that is, a subset of development data to be used in the modelling phase. The choice of attributes is thus essential.⁴ For example, one would expect that patients' age and

⁴ The choice is typically based on cooperation of the producers and users who are experts in the domain.

weight attributes might be important for a disease classification AI system, while the name of the patient has no predictive power.

Subsequently, the attributes are split into training and testing data. The training data is used to build the actual model, for example, a neural network, while the testing data is employed to estimate the model quality. Both data splits undergo labelling, that is, human annotations defining expected outputs with respect to the system inputs needed to train and evaluate the ML models.

Data represent the vein of AI technologies. If they do not reflect reality, including its genuine diversity, accurately, the picture they bring into the AI system is distorted from the outset. Hence, the risk of violating the prohibition of discrimination has its roots in this stage (as explained further in section 3.1). Furthermore, biased data may negatively impact other human rights depending on the domain of deployment (section 3.2). At the same time, the right to privacy might be put at risk, too, if personal data is processed without necessary legal guaranties (section 3.3).

3.1 Prohibition of Discrimination

Fragmented, incomplete, or distorted data—be it intentionally or unconsciously—constitute the primary and fundamental root cause of bias.

Bias based on protected values, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth, or other status,⁵ can be deliberately or unconsciously quite easily introduced into the AI life cycle operating on the basis of the data flow. Data can directly contain such an element or allow for its indirect reflection. If the data is not representative and is marked by a bias, the result produced might be discriminatory. Such ‘a difference in treatment of persons in analogous, or relevantly similar situation’,⁶ or a failure to treat persons differently in relevantly different situations, is unlawful unless it pursues a legitimate aim and the means employed are reasonably proportionate to this aim.⁷ The risk of both direct and indirect discrimination might arise.⁸

⁵ The list of protected values is not exhaustive. See eg art 26 of the International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR); and art 14 of the Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR).

⁶ Specification formulated by the European Court of Human Rights (ECtHR). Eg *Biao v Denmark* App no 38590/10 (ECtHR, 24 May 2016), para 89; *Carson v United Kingdom* App no 42184/05 (ECtHR, 16 March 2010), para 61; the UN Human Rights Committee (UNHRC) refers simply to ‘differentiation of treatment’. See UNHRC ‘General Comment No 18: Non-discrimination’ (37th session, 1989) UN Doc HRI/GEN/1/Rev.9, para 13.

⁷ ‘[A]n objective and reasonable justification’ is required as formulated by the ECtHR. See eg *Molla Sali v Greece* App no 20452/14 (ECtHR, 19 December 2018), para 135; *DH v Czech Republic* App no 57325/00 (ECtHR, 13 November 2007), para 175. See also UNHRC, *ibid*.

⁸ For indirect discrimination, see eg Anya Prince and Daniel Schwarcz, ‘Proxy Discrimination in the Age of Artificial Intelligence and Big Data’ (2020) 105 Iowa Law Review 1257.

This situation may, for example, materialise when the development data provider of an AI system aimed to estimate the probability of recidivism provides incomplete data based on the police and investigation databases of only selected neighbourhoods, or with a focus on one gender only. The output produced by the AI technology might be biased against a specific ethnic or religious minority or gender. The input data will carry an unsubstantiated weight of some patterns (that might include the protected value) that will be reproduced in the AI system. A similar case is a hiring AI system utilised to recruit, test, and predict the success rate of applicants, which is biased, for example, against women. The bias could be caused by using a training data set containing samples selected mainly from the male population, whose representation in that job sector was prevalent in the past years.⁹

Bias can be introduced into AI technology on several levels: via input data,¹⁰ definition of the input attributes, training and testing data split, and data labelling, but also via the selection of metrics, transfer learning from any predecessor model and the deployment setup. It may also arise during the operation of the AI system if AI retraining is used,¹¹ or by inappropriate synthetic data use.¹²

The challenge is how to avoid such biases or how to possibly compensate for them. The primary question is whether the data carrying the protected value can actually be taken into account or whether they must be completely eliminated from the AI system. The elimination of the protected value from the input data, or more precisely from the selected attributes, is in fact not a panacea: there are situations that do require the presence of a specific protected value. Moreover, the elimination is not always technically feasible in practice. The following three main principles apply:

- (i) The protected value *may be taken into account* if the differentiated treatment on its basis is justified and proportionate to the legitimate aim pursued. An example could be the use of AI in the field of dermatology. Here, the criterion of skin colour could be relevant for the precise health diagnosis, and certain physical traits constitute essential features. Similarly, the selection process within a film casting based on gender or age could be justified and proportionate to the aim pursued.¹³ Therefore, the protected

⁹ Amazon faced a similar flaw in their AI recruiting tool in 2018, see Jeffrey Dastin, ‘Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women’ (*Reuters*, 11 October 2018) <www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.

¹⁰ The term ‘input data’ refers in this paper to the data collected both for the development of the AI and during the AI operation. If specifically one subset is concerned, the terms ‘development data’ or ‘operational data’ are used.

¹¹ Although the causes of the insertion of bias and applicable principles are the same, the remedy is different in the phase of deployment, see section 6.2. Malicious use of AI technology for discriminatory objectives is addressed separately in section 6.3.

¹² See section 3.3.

¹³ See, however, eg Aja Romano, ‘The Debate over Bridgerton and Race’ (*Vox*, 7 January 2021) <www.vox.com/22215076/bridgerton-race-racism-historical-accuracy-alternate-history>. This raises the

value may lawfully figure among the attributes in the AI system and does not have to be erased.

- (ii) The protected value *shall not be taken into account* if the differentiated treatment, (dis)advantaging one group or its member, would not be justified and proportionate to the aim. In this scenario, the information about ethnicity or gender shall not figure among the attributes constituting the AI technology predicting recidivism. Similarly, an AI-based bank loan management system shall not take into account gender or indirect factors such as maternity leave.
- (iii) The protected value *shall be taken into account on an equal basis* to prevent the (dis)advantaging of one group or its member. The aim is to ensure objective representation without discrimination. This scenario can be envisaged if the protected value constitutes the indispensable determining element underpinning the very function of the AI technology, for example, facial recognition where the presence of samples of all external facial features, including colour, will be needed in order to prevent the (dis)advantage. Furthermore, such a need may arise if the protected value is contained in the input data and the differentiated treatment based on it is not justified and proportionate, however, for one reason or another, it is not technically possible to erase the protected value. Examples may include voice recognition where the input voice cannot be made neutral, medical expert systems, statistical evaluation for marketing purposes, and so on. Equal representation of the protected value can be ensured, for example, via balancing or weighting the data set, including adjusting for missing data.

Other challenges to the prohibition of discrimination such as indirect discrimination or the open-ended list of protected values merit further research. The underlying elements of uncertainty present in the human world are all the more amplified if translation into computational models is required.

3.2 Multiplication of Human Rights Violations Provoked by Biased Data

If the input data, or the selection of attributes, are biased, the violation of the prohibition of discrimination may be supplemented by the violation of the respective human right in the field of AI deployment. For example, if the entrance gate with facial recognition mechanism at the airport erroneously does not recognise the facial features of persons of a certain ethnicity and denies access to the airplane, it

issue of a continuous update of AI systems in view of the evolution of human rights values, see section 6.5.

violates not only the prohibition of discrimination but also the freedom of movement.¹⁴ Similarly, biased data used in predictive justice and policing operating on the basis of probability estimation of recidivism¹⁵ might cause not only discrimination but, consequently, also the violation of for instance the right to liberty, fair trial rights, or the rights of the child.¹⁶ The right to an adequate standard of living might be violated by a biased AI system indicating the creditworthiness of individuals if used to determine access to housing.¹⁷ In this scenario a multiplication of human rights violations arises: the primary violation resides in the violation of the prohibition of discrimination, and the secondary violation impacts the relevant human right in the particular field of the AI deployment.

While the key to the prevention of human rights violations due to biased data, or unbalanced data in general, resides in the user requirements setting and the general control mechanisms of the AI life cycle, the secondary violation of human right itself might be mitigated, to a certain extent, with respect to the degree of autonomy of the AI system. The extent of the participation of the human factor and the resulting legal accountability for the human rights violation, in other words, the relation between AI technology and the user, is relevant. If the AI technology erroneously qualifies an individual as the criminal suspect under a search warrant and the police adequately intervene and detain the person, the user participates in the secondary violation of a human right (right to liberty). Is there, in a similar vein, a participation of the user (airport), if the entrance gate at the airport based on facial recognition AI system erroneously blocks a passenger and denies access? And what about a fully autonomous vehicle operating on biased data and executing false operative steps causing an accident with great material damage, bodily harm, and loss of life, which impacts on the right to property, right to health, and right to life? Is the element of discretion, or the human factor, present here?

¹⁴ For the impact of a biased facial recognition technology on, for example, unlawful detention by police, see TJ Benedict, ‘The Computer Got It Wrong: Facial Recognition Technology and Establishing Probable Cause to Arrest’ (2022) 79(2) Washington and Lee Law Review 849.

¹⁵ Eg the COMPAS system used in several US state jurisdictions to predict the risk of reoffending, assessed by the investigative website *ProPublica*. See Julia Angwin and others, ‘Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks’ (*ProPublica*, 23 May 2016) <www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. See also Harm Assessment Risk Tool (HART), used by Durham Police in the UK: Marion Oswald and others, ‘Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and “Experimental” Proportionality’ (2018) 27 Information and Communications Technology Law 223.

¹⁶ See eg the ProKid 12-SI that evaluates the risk of the criminality of twelve-year-old children, used by the police in the Netherlands since 2009: Karolina La Fors-Owczynik, ‘Profiling “Anomalies” and the Anomalies of Profiling: Digitalized Risk Assessments of Dutch Youth and the New European Data Protection Regime’ in Samantha Adams, Nadezhda Purtova, and Ronlad Leenes (eds), *Under Observation: The Interplay Between eHealth and Surveillance* (Springer 2017).

¹⁷ Comment on the SCHUFA system in Germany in European Digital Rights (EDRI), ‘Use Cases: Impermissible AI and Fundamental Rights Breaches’ (August 2020) <<https://edri.org/wp-content/uploads/2021/06/Case-studies-Impermissible-AI-biometrics-September-2020.pdf>> 7.

The inclusion of the human element in the deployment of the AI technology might certainly diminish the emergence of the secondary violation of human right if the discretion is used efficiently and the biased AI decision is reviewed and remedied. Even though the stagnation of the AI development at the level of assistive AI systems is not realistic, the prerequisite of the presence of a human factor might be introduced with respect to some critical AI system domains like justice, health-care, and others.¹⁸ The review of the output of the assistive AI technology requires, however, certain guarantees based on informedness and expertise in order for the user to prevent effectively and efficiently the secondary violation of human rights. Transparency of the AI decision, which ensures the necessary information for the user is an important tool in this respect, too.

3.3 Right to Privacy

AI technologies might pose a risk to the right to privacy in particular via the protection of personal data. Data is the red thread of AI and an important volume covers personal data. Personal data comprises any information which is related to an identified or identifiable natural person such as name, identification card number, or biometric data. Personal data can be present in both input data and as processed information¹⁹ and is used across various technologies, be they based on facial recognition, medical diagnostics, geographical location, or purchasing preferences. The right to privacy is potentially at risk during the data collection and data processing both in the development and deployment²⁰ phases of the AI technology.

Particular questions arise in practice: does the AI engineer work with anonymised data or is pseudonymisation used? Is it feasible to substitute the 'real' input data with synthetic data? In which domains and under what conditions is the use of synthetic data advisable?

Anonymisation is used notably in healthcare, mobile device services, image processing, census, data monetisation, reporting to third parties, and other fields. It is effectuated on the level of removing personally identifiable information from the data set. Anonymisation still does not, however, ensure the full-fledged guarantee of non-reconstruction of the original data.

Pseudonymisation, on the contrary, is used in the way that the information that can point to the identity of a subject is replaced by 'pseudonyms' or identifiers

¹⁸ See, however, Zana Buçinca, Maja Malaya, and Krzysztof Gajos, 'To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-Making' (2021) 5 Proceedings of the ACM on Human-Computer Interaction 1.

¹⁹ Processed information involves any form of input data other than raw, eg data aggregation or parameters of a trained ML model.

²⁰ For the additional issues of the deployment phase, see section 6.2.

which prevents the data from specifically pinpointing the subject but still enabling the latter's identification. Pseudonymisation is used, for example, for employee reviews, anonymous statistical evaluations, tracking customer spending data, analysing users for targeted advertising, and others.

Pseudonymisation preserves the individualisation of the analysed data set while anonymisation breaks the bindings (eg erases the tracking activities such as web browser cookies). An example could be the AI system in a hospital assisting a kidney transplant committee in its decision of approval. The data of the patients submitted to the selection process for the transplantation have to be pseudonymised in order to maintain the distinction among the patients without knowing their identity to avoid bias. The anonymisation is not appropriate since the personally identifiable information would be removed from the data without replacement and, accordingly, the binding to a particular person would be broken, albeit the identity is needed in order to apply the decision of approval.

The use of synthetic data is feasible, although to a rather restrictive extent. In fact, it runs counter to the gist of the AI which processes reality, all the more so in case of the analysis of big data. Replacing the real data by artificial samples can significantly harm the representativeness of the data set. At the same time, if it is possible to generate synthetic data, which would fully replace the real data, the algorithmic solution to the problem is often known and, thus, there is no need for ML. In practice, synthetic data is used to supplement development data where it is not possible to cover the entire set of examples for time or practical reasons. For example, the Tesla car company uses synthetic data to train situations where physical data from real traffic is missing, or their collection would be too time-consuming. Another example is Amazon, using synthetic data to train Alexa's language locales like Hindi, US-Spanish, or Brazilian Portuguese.²¹

In order to avoid the risk of violating the right to privacy when dealing with personal data, the following basic approaches should, in general, be highlighted with respect to AI: first, a preference for closed systems, meaning those that do not provide information to the third parties and utilise it only for the declared objective,²² second, the anonymisation or pseudonymisation of the input data; and third, the consent given by the users to the processing of their private data and further use.

²¹ See also Janet Slifka, 'Tools for Generating Synthetic Data Helped Bootstrap Alexa's New-language Releases' (*Amazon Science*, 11 October 2019) <www.amazon.science/blog/tools-for-generating-synthetic-data-helped-bootstrap-alexa-s-new-language-releases>.

²² For example that the Alexa-style voice assistants use the data only for the self-training and do not transfer them to the third parties or that social media do not provide the data to the third parties—see for instance, the 2010 Cambridge Analytica affair. See eg Alexander J Wulf and Ognyan Seizov, ‘“Please Understand We Cannot Provide Further Information”: Evaluating Content and Transparency of GDPR-mandated AI Disclosures’ (2022) *AI & Society* <<https://doi.org/10.1007/s00146-022-01424-z>>.

4 Modelling

The modelling phase involves the selection of an appropriate model type and related algorithms of the training pipeline as well as the actual model training. The training is typically performed using the development training data only. The testing data in conjunction with an evaluation metric are then used to assess the model's operational performance. The impairment of human rights may arise from evaluation metrics and transfer learning (section 4.1) and unexplainable models (section 4.2).

4.1 Evaluation Metrics and Transfer Learning

The selection of metrics and transfer learning risk introducing bias into the AI technology. Evaluation metrics are vital indicators of AI system quality. Typically, there are multiple metrics involved. For example, the AI engineer may decide on different metrics in the modelling, evaluation, and deployment stages. These metrics can span from low-level approaches such as classification accuracy (the ratio of correctly classified samples of testing data, eg images) to complex ones (eg a projected profit of an automated stock market bidding system). The metrics can potentially introduce bias into the overall system. For example, when unbalanced data is involved, the wrong performance metric selection can prevent the AI engineer from detecting possible problems within sparsely covered data categories. As a result, predictions of such a model will be more or less random for less represented minorities.

Another potential problem comes from the widespread reuse of pre-trained models. This approach is known as transfer learning, and it allows training highly complex ML models even when the training data is limited in size. As an example, most state of the art natural language processing (NLP) methods, dealing with tasks such as text classification or language translation are based on fine-tuning an existing pre-trained language model. This language model would be typically supplied by a third party, which makes it hard or impossible to access the original training data or even the details of the training procedure. The potential problem then lies in the unintended propagation of the language model biases to the target model.

4.2 Unexplainable Models

Unexplainable AI behaves as a black box which refers to models that are too complex to be straightforwardly interpretable by humans. This involves notably AI based on ML techniques that train models of such complexity like neural

networks, a widely used paradigm today. Other model types, for example, decision trees, are close to human natural reasoning and inherently enable a higher degree of explainability. Nearest neighbour models supply explanations based on the analogy. Similarly, graphical models give an idea of observed events' probabilistic dependencies.²³ In principle, only the simplest models, such as small decision trees, can provide full-fledged explanations of the predictions.

The fundamental human right which is particularly at stake today with respect to an unexplainable AI is the right to a fair trial. The right to a fair trial, or due process rights, constitute the fundamental procedural right that assures the potential victim of a human right violation the access to an independent and impartial court along with all necessary legal and procedural guarantees.²⁴ The United Nations Human Rights Committee (UNHRC) in its General Comment No 32 highlights:

The right to equality before courts and tribunals also ensures equality of arms ... The principle of equality between parties applies also to civil proceedings, and demands, inter alia, that each side be given the opportunity to contest all the arguments and evidence adduced by the other party.²⁵

Furthermore, the European Court of Human Rights (ECtHR) notes in the 'Guide on Article 6 of the European Convention on Human Rights—Right to a Fair Trial' that '[n]on-disclosure of evidence to the defence may breach equality of arms as well as the right to an adversarial hearing'.²⁶

In practice, the producer of a black box AI system is not capable of rendering an explanation as to how the system arrived at the selection of the information it based its decision upon. The lack of this information undermines or paralyses the victim's defence. If a complex AI-based system profiling individuals determines the access to or delivery of public services such as social security, policing or migration control, the denial of such evidence thwarts the effective and efficient exercise of the victim's fair trial rights. Similarly, if an employee is denied equal opportunity to be promoted to an appropriate higher level based on an AI screening, the absence of access to the information supporting such a decision undermines the person's possible defence. Making AI systems reveal their internal logic thus constitutes the

²³ For more information about the explainability of particular AI models, see eg Christoph Molnar, 'Interpretable Machine Learning: A Guide for Making Black Box Models Explainable' (29 March 2022) <<https://christophm.github.io/interpretable-ml-book/>>.

²⁴ Eg ICCPR art 14.

²⁵ United Nations Human Rights Committee (UNHRC), 'General Comment No 32, Article 14: Right to Equality Before Courts and Tribunals and to a Fair Trial' (23 August 2007) UN Doc CCPR/C/GC/32, para 13.

²⁶ European Court of Human Rights (ECtHR), 'Guide on Article 6 of the European Convention on Human Rights—Right to a Fair Trial (Criminal Limb)' (30 April 2022) <www.echr.coe.int/documents/guide_art_6_criminal_eng.pdf>, para 174. See also ECtHR, 'Guide on Article 6 of the European Convention on Human Rights—Right to a Fair Trial (Civil Limb)' (31 December 2021) <www.echr.coe.int/documents/guide_art_6_eng.pdf>, para 377.

cornerstone of the rule of law and fair trial rights. AI shall be able to justify its decisions and provide the needed evidence, in the same vein as humans are obliged.

How to overcome this seemingly unbeatable challenge? One option is to support the alternative methods of mitigation of unexplainability (section 4.2.1) while the other is to restrict the use of a black box (section 4.2.2).

4.2.1 Mitigation Methods for Unexplainability

Efforts to mitigate the impacts of unexplainable AI could take two basic forms: a creation of a surrogate model aiming to explain the black box, and ex post review methods based on a counterfactual.

A common method to help explain an AI system is to train a surrogate model to mimic the black box model. Such an explanatory model constitutes a simplified copy model for a domain expert (eg a doctor with respect to health diagnostics or a judge regarding predictive justice) or an accessible visualisation. Although it can detect and identify the most important root causes of the AI decision and its core features, the deficit resides in a lower expressiveness and the loss of information of the surrogate. Moreover, AI systems are highly complex, thus—without the background knowledge—their efficient translation would be difficult. The field of explainable AI is currently subjected to extensive research.²⁷ The large part of these activities focuses on explanation methods for neural network models.

Alternatively, an ex post comparison method providing information about the functioning of AI technology is based on a counterfactual. The developer of the AI system could provide a test bed—an interface to enable the third party, or more precisely the claimant, to verify the properties of the AI technology in other cases. The system could give an example of what input data would be needed to alter the AI prediction in a particular case. For instance, if an AI-based financial service denies a loan to the claimant, the counterfactual might indicate the threshold at which the loan would be approved, for instance, the minimum salary required for eligibility. As another example, the counterfactual approach may help to easily demonstrate that a predictive justice AI system generates decisions based on attributes that cannot be influenced by the claimant (like their birthplace) and, as such, reveal even forms of indirect discrimination. If the testing of the predictions is effectuated directly by the producer, the objectivity is, nevertheless, not assured with respect to the exclusive and targeted high-tech knowledge of the AI system.

4.2.2 Restricting the Use of Black Box Systems

Another solution is to restrict or ban black box systems, at least in critical AI applications, and shift to the white box algorithms. The latter are AI systems that

²⁷ See eg Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić, 'Explainable Artificial Intelligence: A Survey' (2018) 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) 210.

are explainable but, in most cases, significantly less powerful than their black box counterparts. Notably, for analysis of very big data, such as a database of high-resolution images or textual databases, a black box system is most likely indispensable. The white box models are often deployed in life-critical scenarios such as healthcare. They are mostly small instances of models based on specific paradigms allowing for interpretable visualisation.²⁸

Similarly to the existing concepts of ‘life-critical’ systems (eg pharmaceutical ingredient dosage estimations, healthcare systems) and ‘mission-critical’ systems (eg airplanes, cars, and other transportation means), which are subjected to stricter requirements in the process of approval for operation, an idea of ‘human rights-critical’ AI could be introduced. The ‘human rights-critical’ AI technologies marked by a high risk of violation of human rights with severe impact on the human would at least always need to be certified. Their list is to be appropriately determined while it is conceivable to encompass, for example, predictive justice, robotic surgery, or autonomous weapon systems (AWS).

By explainability is meant algorithmic explainability, that is, the clarification of the manner through which a decision is made in a way that it is understandable to a human. This is necessary, for example, in a number of life and mission-critical applications, where full control over all parameters and procedures which lead to the output of the system during system design is needed. Examples include pacemakers, aircraft control systems, and others. Explainability needs to be distinguished from transparency which includes development and operational transparency. The latter reveals the information the AI decision has been based upon, in other words, which input data and with what weight have been taken into account.

Moreover, these concepts shall be distinguished from confidence and accuracy, which serve another goal. Both express the level of precision of the AI technology: while accuracy refers to the quality of the AI system and indicates, for example, the estimated percentage of cases in which its predictions will be correct,²⁹ confidence expresses the subjective evaluation of the result by the AI technology itself. In the latter case, for example, in the domain of image recognition the AI technology indicates that the scrutinised image amounts by 60 per cent to a dog and by 40 per cent to a cat and provides such an auto-assessment of precision.

It is important to know that unexplainability does not always hinder the functioning of the AI system from the perspective of human rights. In practice, depending on the type of AI technology determined by the margin of decision-making of the user, diverse requirements mentioned above are relevant. In other words, if a human control is involved in the operation of the AI, the requirements

²⁸ An example includes decision trees, nearest neighbours, linear/logistic regression, naïve Bayes classifiers, or custom mathematical models.

²⁹ This ‘classification accuracy’ informs about the percentage of cases in which the AI technology provided a correct response (ie the proportion of correct predictions based on testing data) and thus estimates how the AI technology will behave in the real operation.

change and the criterion of explainability might yield to other requirements. At one end of the spectrum, the assistive AI technologies leave the final decision in the hands of the user. They include assisted and augmented AI. Assisted AI uses the combined power of data, cloud, and data science to help people make decisions, leaving the final decision in the hands of the human users. While it improves things that humans are capable of doing, it requires continuous human input and intervention. Augmented intelligence, on the other hand, is a human-centred co-operation of people and AI enabling humans to do things they could not otherwise do, thanks to the informational extension of their perception, task understanding, decision-making, and creating new experiences.

A user benefiting from the aid of an assistive AI will profit not from the algorithmic explainability but the transparency of the AI decision, in order to support their decision and verify the credibility and efficiency of the AI output. The knowledge of what input information contributed to the AI system predictions and with what weight serves the user as the proof of the reliability of the AI system in order to adopt the final decision. This will be notably required in an expert domain based on professional qualifications, such as healthcare, law, cybersecurity systems, and others. For a judge sentencing the accused to imprisonment, for example, and taking into account the AI output regarding the degree of probability of recidivism, the algorithmic explainability would not serve much purpose. What is needed here is the transparency of the AI decision, meaning the information about which input data or criteria the AI solution is based on. Algorithmic explainability is important at the level of the system development, but it does not convey additional information to the user to support their decision-making, even in the case of algorithmically explainable systems. On the contrary, a general public using an AI-based smartphone suggesting intuitive writing of messages benefits neither from explainability, nor transparency of the AI system, as the correctness of the result can be deduced directly from the suggested text. On a practical level, if high-tech expertise is needed to understand algorithmic explainability, certain AI expertise is still expected also for the understanding of AI transparency. It is presumed therefore that AI support will be indispensable, similarly to today's commonplace IT assistance, in various fields of activity benefiting from AI solutions.

By contrast, at the other end of the spectrum stands the most advanced form of AI, autonomous (or automated) intelligence, that can perform a complete action, including information gathering and decision-making, without a human control or intervention. For the moment, while the number of autonomous intelligence applications is increasing, handing the total control over to machines is still a question. Besides the issues of accountability, this is especially due to the complex tasks or diverse input data where a variety of different results is possible in combination with fluctuation of input data. The rising degree of autonomy of decision suppresses the requirement of an a priori operational transparency while the high degree of accuracy is essential for the user. The user needs an autonomous AI

technology that is reliable, and thus the requirement might be set to a high degree of accuracy, close to 100 per cent. The operational transparency is, however, indispensable ex post if due process rights are at stake.

To sum up, revealing the internal processes of an AI technology and providing access to information provided to the claimant, are the prerequisites of the effective enjoyment of the right to fair trial. Explainability and transparency, as well as accuracy, are essential factors in this respect. Although the algorithmic explainability will not have, in practice, much informative capability for the user and would not, thus, be effectively useful, the criterion of explainability may serve as the requirement for the certification of the high-risk AI technology—labelled ‘human rights-critical’. It could condition its deployment and, consequently, eliminate the black box AI. The criterion of transparency may be required for all AI technologies at a different time—for the subject matter experts using an assistive AI technology before they take the final decision and for all autonomous AI technologies for the sake of ex post review if requested.

The transparency of decision, however, does not overcome the existence of a black box. Transparency only partially helps to reveal the information about its internal processes and, thus, supports fair trial rights. Moreover, some questions remain open for discussion: In order to fully ensure the right to a fair trial, is explainability as the requirement for the certification of *all* ‘human rights-critical’ AI the panacea? Which AI technology is actually not ‘human rights-critical’?

5 Evaluation

In this phase, the system is tested and verified to ensure that it meets the business objectives. The evaluation is based on one or more metrics established to assess selected properties (qualities) of the model. The testing and verification help to detect the possible source of human rights violation.

During the evaluation phase, no new risk to human rights is inserted, however, a challenge might arise that an already existing risk of a human right violation, having its roots in previous stages of the AI life cycle, will not be detected. This might be the case if inappropriate evaluation metrics are selected or the testing data is not representative.

The AI evaluation is a part of a standard software development process. It should always follow the specifications of the AI system. Therefore, in order for the human rights risk assessment to be involved in the evaluation, the criteria should be already covered by the system specifications and requirements. The evaluation is done during the development phase, right before deployment, and can be further extended to system monitoring during the deployment. The evaluation consists of verification (specifications are correctly implemented), validation (system works as the user requested), and qualification (system complies to regulations, standards, and certifications).

6 Deployment

The deployment phase sets out when the system becomes fully operational for the user. Risks to human rights may arise typically in five basic situations: deployment of the AI system not matching the target operating environment (section 6.1); data processing during operation (section 6.2); malicious use of AI technology (section 6.3); abuse of AI technology (section 6.4); and non-transparent operation of the AI system (analysed under the general question of the transparency of AI system in section 7.2). Monitoring of the AI system thus constitutes an essential element of the AI life cycle in order to ensure its correct functioning as well as its compliance with human rights norms (section 6.5).

6.1 Deployment of AI System Not Matching the Target Operating Environment

Human rights violation may be provoked by technical shortcomings during the deployment of the AI system, such as operation in a different context than for which the technology was originally designed or due to insufficiently identified system boundary conditions.

In the first case, the AI technology has been trained for a particular context—be it geographical, cultural, social, etc—but it is deployed in another environment. For example, an AI-based school exam evaluation system has been trained on data from the majority of schools in a country not supporting inclusive education—however, it was applied in an educational institution with students with special needs. The results would be discriminatory and the right to education might be impacted. Similarly, if an AV has been developed for the automotive market in the United States (US), it will mark substantial deficiencies if deployed in a country, for example, with Arabic script or with different traffic signs like in Asia or Europe.³⁰ The different geographical deployment might put several human rights at risks.

As concerns the boundary conditions, the AI technology functions correctly; however, some development data is lacking and, thus, at certain conditions, the system does not provide accurate results. For example, the AV is trained for all sorts of weather conditions in the US, yet the drizzle or roads under extreme snow have not been anticipated for visual object recognition. In these unexpected conditions, the malfunctioning of the AI technology may again endanger human life or cause other harm.

What is the remedy to such unanticipated situations leading to the violation of human rights? Similar to the prevention of other risks to human rights, the core is the dialogue between the producer and customer, here precisely the user. The

³⁰ See eg Abdul Azeem Sikander and Hamza Ali, 'Image Classification Using CNN for Traffic Signs in Pakistan' (2021) <<https://doi.org/10.48550/arXiv.2102.10130>>.

context of the deployment and boundary conditions of the AI system should be clearly set in the specifications. For example, that the AV is intended for traffic deployment in the US. Moreover, the context and boundary conditions should be verified during the process of approval for the operation of the AI technology.

If the context is fixed in the specifications, it will be taken into account and verified during the testing stage as it screens the AI technology against the initial requirements. Boundary conditions, however, will often be, as unpredicted actual facts, detected only during the operation of the AI technology. In other words, ex post, when the problem has already arisen and entailed the human rights violation. For example, the system for determining social benefits was not prepared for borderline emergencies such as natural disasters and continues to apply the standard benefits for normal situations. Although the AI technology might seem to function correctly, failure to confirm boundary conditions can arise in decisions that are not adjusted for measured or detected extremes or rare situations. In this case, even the monitoring does not prevent the human rights violation. Nevertheless, the monitoring still remains essential in order to detect the error, to adopt the appropriate remedy, and to prevent the repetition of the human rights violation.

6.2 Data Processing During Operation

The AI system operates based on operational input data. This resides in the data processing and evaluation and it could also include the improvement of the AI system functionality.

A particular question arises in this respect: Is it possible to restrict the type of input data and their processing mode, which the AI technology will be collecting and processing itself during its deployment into operation? For example, to restrict Alexa not to listen to, store, and process personal or sensitive information that is not addressed to her but, at the same time, may be uttered in the household the device is placed in? The restriction of the type of data ‘absorbed’ by the AI technology in operation in order to protect the right to privacy is a tricky one. This issue was particularly relevant, albeit on the business level, with respect to the COVID era home office phenomenon when several companies instructed their employees to switch off Alexa and other home AI appliances to avoid business information leakage and eavesdropping.³¹ Not only cybersecurity is at stake; it should be borne in mind that even the simple processing of information by Alexa contains the element of data storage in the cloud. The user might not be aware of this technical data channelling

³¹ See eg Karen Sloan, ‘Calif. Bar to Attorneys: Disable Alexa when Working from Home’ (*Reuters*, 13 August 2021) <www.reuters.com/legal/legalindustry/calif-bar-attorneys-disable-alexa-when-working-home-2021-08-13/>.

or the AI appliance might ‘wake up’ at random. The challenge is how to set up the AI technology *in advance*, during the ML process, to switch off if personal or other sensitive information is at stake, which is hardly imaginable before the user actually expresses them. Since the data processing—and thus analysis—is performed in the cloud, the data entry into this cloud processing—in other words, the ‘data absorption’ phase—is unavoidable. The only restriction conceivable is the immediate elimination of the data from the cloud after the qualification. One cannot guarantee, however, that the data log will not be retained. The protection of personal data in this particular scenario does not reside, therefore, primarily in the technical field but in the human approach and preventive behaviour.

As concerns the AI system improvement, or fine-tuning, based on the operational data, this approach is known as online learning and it is common for highly dynamic domains. For example, we expect the system to swiftly react to an introduction of new kinds of goods, such as an e-shop recommender system that continually collects buyer feedback and periodically fine-tunes its models. During the AI system improvement, the same risks that appear in the data preparation phase might arise with respect to the introduction of a bias (and unbalanced data in general) and the protection of the right to privacy.

Although the root causes of the introduction of bias and applicable principles analysed in section 3.1) remain the same, the solution to the risk of unjustified discrimination is different in the deployment stage. This is especially true if the functionality of the system is adjusted automatically during operation without a human intervention. The system should be designed in a way that its functionality can only be changed within predefined limits so that new information does not overwrite the original settings (eg that body tan is not associated with colour). Operational data should be automatically filtered and faulty information detection should be implemented in order to avoid an external manipulation of the system during operation. For example, an autonomous driving system with sign recognition could be manipulated to recognise signs inserted in the false context. As concerns the mission-critical systems, such as autonomous driving, the retraining of key functionalities is therefore not performed in operation; the system sends the data to the manufacturer who performs the retraining under controlled conditions.

Similarly, the protection of privacy with respect to personal data or sensitive information faces different challenges during the deployment. If the improvement of system functionality is done internally, the misuse of personal and sensitive data is low as it is not stored after processing and is thus forgotten. On the contrary, if the data is sent to the developer during the retraining (like in case of speech recognition systems working in the cloud, including home appliances), the risk to the right to privacy arises.

6.3 Malicious Use of AI Technology

Another situation where AI technology might impact on human rights is when it is used for an illegal objective. Unlawful monitoring of citizens violating their freedom of movement or right to privacy is already utilised by certain governments.³² So-called killing drones and other lethal autonomous weapon systems (LAWS) endanger human lives.³³ Password cracking might be used for malicious objectives. Deepfakes produced by AI can manipulate the user, for example, via generated clickbait headlines and distorted images and videos.

It is surely possible to adopt appropriate legislation that prohibits the development of certain AI technologies or their use in a particular field (eg manipulative techniques beyond a person's consciousness distorting a person's behaviour and causing harm or social crediting system leading to unjustified discrimination)³⁴ or legislation that subjects the AI development and operation to legal restrictions or licensing (similarly to a firearm licence, licence for hazardous activities, and so on). Exceptions could apply to specific actors like state security agents and others (eg the use of biometric tools in public spaces restricted only to law enforcement and under certain conditions).

On a practical level, the question arises whether there exists a technical measure that could effectively prevent the malicious use of an AI technology and, thus, human rights violations. A technical switch that would stop or turn off the AI technology if used for other objectives than specified in the business requirements.

The technical domain offers standard IT processes and the AI does not present any particularity in this respect, similarly to the question of legislation. The guarantee is twofold: first, the supplier verifies to whom the system is delivered and, second, both the supplier and the operator maintain the capacity to turn off or

³² For the overview of the deployment by states of AI surveillance tools to monitor, track, and surveil citizens, see the *AI Global Surveillance (AIGS) Index*. Carnegie Endowment for International Peace and Steven Feldstein, 'The Global Expansion of AI Surveillance' (*Carnegie Endowment for International Peace*, 17 September 2019) <<https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847>>.

³³ See eg UN Security Council (UNSC), 'Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011)' (8 March 2021) UN Doc S/2021/229, para 63. For a legal assessment of the use of LAWS, see eg Elliot Winter, 'The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law' (2022) 27 Journal of Conflict and Security Law 1; Jai Galliott, Duncan MacIntosh, and Jens David Ohlin (eds), *Lethal Autonomous Weapons: Re-examining the Law and Ethics of Robotic Warfare* (OUP 2021); and Andrea Spagnolo, 'What do Human Rights Really Say About the Use of Autonomous Weapons Systems for Law Enforcement Purposes?' in Elena Carpanelli and Nicole Lazzerini (eds), *Use and Misuse of New Technologies: Contemporary Challenges in International and European Law* (Springer 2019) 55.

³⁴ See eg the list of 'Prohibited artificial intelligence practices' laid down in art 5 of Title II of the European Commission's 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>. Although not expressly linked to human rights, all the indicated practices relate to human rights violations.

restrict the operation of the AI technology. The first option is effectuated in a common manner via licences that are verified at the start of the use of the system (password, subscription, and standard remote control techniques). For example, the use of an AI-based technology operating with medical data will be activated only by a licensed subscriber such as a specific hospital or a doctor as the user, in order to protect the right to privacy. The monitoring of IP addresses, too, serves this purpose. The second option, the paralysing interference into the operation of the AI system, can be performed either at the distance, if irregularities are detected, or an automatic operational restriction mechanism can be integrated into the system. For example, a camera surveillance system must periodically log on to the server to continue recognising people under the set conditions. The geolocation restrictions, for instance, can be introduced by the producer even on an ad hoc basis. If it is revealed that some state uses the AI technology for malicious objectives violating human rights, for example, social crediting, the supplier or operator might ban the access for that state. Typically, intelligence service operates AI devices with built-in remote operation controls, including activation and deactivation.

All AI life cycle actors involved and the state have variable interests to have such guarantees of control included. Depending on the actor, such guarantee is set in the requirements and the state regulation.

Finally, concerns about the malicious use of AI reflect sometimes the ‘futuristic’ or ‘sci-fi’ vision of an AI that will surmount human control and turn against its human operators.³⁵ Again, it is important to note that the contemporary state of AI technologies, or presumably AI of the near future, does not present any particular risks of malicious use different from those that can be observed in any other IT system. No AI surpasses today the mechanisms which a human being would use to divert the AI operation. It might be faster, but the infrastructure, as well as the mechanisms utilised to overcome the obstacles, are the same as if used by a human or another IT system. A possible guarantee in IT systems today is the restriction or stopping of their operation by an independent system that is integrated into these mechanisms. Examples include critical infrastructures such as the nuclear power plant where the security mechanism is operating independently and is secured at multiple levels and, therefore, is hard to be manipulated by the main operating system itself. The same applies to AI technology.

³⁵ See eg Lance Eliot, ‘Using a Kill-Switch or Red Stop Button for AI is a Dicey Proposition, Including for Self-Driving Cars’ (*Forbes*, 21 October 2020) <www.forbes.com/sites/lanceeliot/2020/10/21/using-a-kill-switch-or-red-stop-button-for-ai-is-a-dicey-proposition-including-for-self-driving-cars/>; and Stuart Russell, ‘How to Stop Superhuman AI Before It Stops Us’ *The New York Times* (8 October 2019) <www.nytimes.com/2019/10/08/opinion/artificial-intelligence.html>.

6.4 Abuse of AI Technology

A particular scenario of a human rights violation produced by the use of AI technology stems from the latter's abuse. An AI technology that is developed in compliance with law and deployed for a legal objective may be attacked and abused. For example, an IoT (Internet of Things) appliance such as voice-controlled home shopping devices and smart home control systems might be hacked and serve as an eavesdropping device violating the right to privacy. An AI-based system for the selection of job candidates might be digitally falsified, which would disadvantage certain groups of candidates interfering with their right to work, or, conversely, it may promote fraudulent candidates to state services. Such potential for abuse concerns almost any field of deployment of the AI technology.

These cases rank to the standard field of criminal law and cybersecurity and, as such, are beyond the scope of this chapter.³⁶

6.5 Monitoring

An important element of the deployment phase with effect on human rights protection is the monitoring of the AI system. The monitoring may detect any discrepancies, deficiencies, and errors that might arise during the operation, including the risks to human rights.

The AI monitoring may be automated or provided by a human, and it can be performed at time intervals specific to the particular domain of deployment. Monitoring means evaluation of the system outputs with respect to the 'expected ones', where the latter can be defined from approved historical results, via anomaly detection system or through a human supervisory, or it may reside in a simple recording of the system outputs. Audit is an evaluation of the whole system, including of the monitoring and operational environment and conditions, and can be formalised. The New York City Council in December 2021, for example, passed the first law in the US banning city employers from using AI-based tools to hire and promote unless those systems can survive an annual audit proving they do not discriminate on the basis of gender or race.³⁷ The audit also serves as important tool for a continuous update of AI systems and human rights risk assessment mechanisms in light of evolving human rights values and their application.

³⁶ And, as such, are beyond the scope of this chapter.

³⁷ New York City Council, Automated Employment Decision Tools, Law no 2021/144 (11 December 2021).

7 Requirements Common to the Whole AI Life Cycle

Although each phase of the AI life cycle reveals specific risks of human rights violations and tailor-made remedies, certain elements aiming to prevent such risks are common to the whole AI life cycle. First and foremost, it is essential to introduce a human rights risk assessment as an integral part of AI systems (section 7.1). Similarly, the requirement of transparency should permeate the entire AI life cycle (section 7.2).

7.1 Human Rights Risk Assessment

In order to ensure the compliance of AI systems with international human rights norms, a human rights risk assessment shall be introduced into the AI life cycle. In other words, it is necessary to integrate the human rights element into the input analysis, requirements setting, and the general control mechanisms.

As noted above, the three core tools setting the concept of the AI system during the business understanding stage are the input analysis and user requirements which then reflect into AI specifications. It is in these conceptual documents that the human rights restrictions shall be set. This ensures that the AI system will be developed, tested, and monitored with regard to human rights limitations.

The actors of the AI life cycle set the requirements depending on the degree of severity of the AI system domain and use. The more critical the domain, the higher the parameters set. The state regulator is expected to set the strictest requirements combining the elements of monitoring, transparency, and explainability in the domain of the ‘human-rights critical’ AI, similar to how it regulates existing life- and mission-critical systems. The lowest level is ensured by the producer who simply reflects the feedback by the users, it means the control is ex post. This applies with respect to AI systems that do not impact human rights or where the risk is low, therefore preventative requirements are not for the moment envisaged. An example is a worsened user experience with respect to a prolonged response time of an application in a smartphone.

A large part of AI technologies will operate based on the conditions set by the customer, which raises the attention to these actors of the AI life cycle and their awareness of the human rights perspective. The customer defines the needs in order for the AI system to work efficiently for the set objective while the human rights aspect shall be included. The dialogue of the AI producer with the customer and regulator is thus central in this respect. For example, judges should be aware of the risk rate of relying solely on the AI-based predictive justice and should have at their disposal all necessary information to review the AI decision. A similar logic applies to personal loan departments of a bank evaluating the financial solvency of an applicant.

The customer should also define, for example, the degree of accuracy of the AI technology in order for the user to gain a real perspective of reliance on the AI decision and the need of the review before the AI decision is implemented in practice. For example, in the domain of health diagnostics, it is a standard today that the customer specifies the degree of accuracy to 100 per cent. On the contrary, for example, within traffic surveillance systems, stock market prediction, text search, there are cases wherein the degree that is required is at a high level but does not need to be 100 per cent. For instance, the use of object recognition software to scan and identify the licence plates for the automated parking control might be set for a lower level, as no critical interest or protected value is at stake. The choice is defined by the criticalness of the AI technology, the field of deployment, the price, and the time of the AI development.

Today, it is not habitual yet for the AI system development and deployment actors to define such requirements. An emphasis should be laid on raising public awareness in this domain. Moreover, it is likely that a judicial institution or employment agency would be capable to define the user requirements (eg the rate of accuracy or the need of transparency) but not the system specifications on AI technology (like input data set coverage and variations, test requirements for all input data, and so on). If a particular domain is not subject to a regulator, it might be practical to introduce some form of AI consultancy or an AI advisory body that would help to define the appropriate criteria for the human rights risk assessment and incorporate them into the requirements. The high technical complexity of AI technologies requires the introduction of an AI expert assistance to be introduced in public and business spheres, as it is common today with generally available IT services.

The compliance with the requirements and system specifications is controlled, *inter alia*, during the approval of the AI system for operation which serves, too, as a safeguard against the human rights violation. The entry of the AI technology into service requires the standard approval of an IT system for operation, including possible certification of the system. It should contain additional AI-specific requirements. The approval of the system for operation can be done at several levels with an increasing binding effect: approval by the producer, approval by the supplier, approval by the customer, approval by an independent private body and, at the highest level, the legally binding approval by the state regulatory authority. The more human rights-critical domain of deployment, the higher degree of authority of approval. The requirements related to the specifics of AI systems should include, in particular, the introduction of continuous monitoring of the deployed system, its transparency, and explainability. Other requirements may be related to the selection of development data and to possible constraints on the system online learning. Depending on the particular AI technology and the conditions of its use, these requirements can be variably combined. In view of the potential human rights impact of AI technologies, the human rights risk assessment elements

should be introduced and extend these requirements. It would verify and test the absence of risks for the human rights violation such as the unjustified presence of the protected value in the attributes or unbalanced data.

Considering the human rights risks posed by AI technologies, the implementation of the above-mentioned procedural and technical restraints in order to prevent violation of human rights is essential. It is important to realise, however, that AI technology is not 100 per cent perfect, in the same way as human decision-making is not. The gist is not to ensure certitude but control. The challenge is to identify and treat the potential risks and root causes, and to introduce the failsafe mechanisms to avoid or diminish human rights violations.

7.2 Transparency

Transparency represents the condition of the credibility of the AI system, its possible revisability and control, and the system compliance with human rights, notably the right to a fair trial. Transparency of the AI system consists of development and operational transparency. The development transparency mainly serves the actors who approve the system for operation. It allows for a retrospective determination whether the system has been developed in a controlled manner and whether all quality requirements have been met. This means checking the continuity of development steps and activities, the link between requirements and their testing, records of revisions and their authors, and so on. Operational transparency, on the other hand, refers to the property of the AI system to provide information on which input data contributed to the outputs of the system and with what weight. In addition, the system supplies information for its monitoring and thus allows retrospective review and audit of its functionality and of the relevance of its outputs. From the human rights perspective, this is essential for the enjoyment of the right to a fair trial. As noted above, the right to a fair trial covers the access to and the right to contest all the arguments and evidence presented by the other party. It is therefore essential that the potential victim of a human rights violation has access to the maximum available information about the functioning of the AI system.

The developer of AI technology should be capable of providing the information about all phases of the development process. The operator should supply the operational records. The main problem with making such information available to third parties is fourfold. First, some of the input data might be confidential. This legal protection might be, however, lifted by a court decision if needed. Second, transfer learning utilised frequently today in ML applications hinders the access to the information about the transferred model, unless, obviously, a legal rule is established that any such model must be accompanied by corresponding information in line with the standardised description of the AI life cycle. Third, rearranging the state or the outputs of the AI system into an intelligible and accessible form and the

adaptation for a particular use-case may be computationally demanding. The data preparation for such reporting is arduous since the system would be likely shaped and understandable for expert actors only. This concerns, too, the establishment of an ex post information and testing mechanism provided to third parties. Does an adequate regulation of the obligation to provide access to information about the AI system have to take into account the financial aspects and so does the judge issuing such an order? Fourth, the intellectual property rights will be at stake and shall be duly considered in view of the often exclusive, high-tech expertise of few.

Particular attention should be paid to the disclosure of information on AI system algorithms used to perform public tasks, notably if the access to or delivery of essential public services is at stake or if enforcement measures are involved. For instance, ‘the algorithm describing verbally or graphically “step by step” the operation’ of the AI-based System of Random Allocation of Cases in Poland, which—since 2018—has decided how to assign cases randomly to judges in common courts, was finally confirmed as *public information* by the Supreme Administrative Court in Poland, and thus mandatorily disclosed.³⁸ Another important step forward is the concept of the so-called AI register, a tool to support the transparency and public control of governmental AI. The City of Amsterdam Algorithm Register³⁹ and the City of Helsinki AI Register⁴⁰ rank among the first public AI registers and provide the following information regarding their AI systems such as the automated parking control or the health centre chatbot: purpose and impacts; accountability; data sets; data processing; non-discrimination; human oversight; risks and their mitigation measures; and explainability. This enhances the trust in public service and access to information essential for the protection of human rights, including due process rights.

8 Conclusion

Artificial intelligence provides a great service to society, nevertheless, it is not just a friend but might be also a foe, to the extent the society allows it. This, after all, applies to all technologies—almost any technology can be misused and become a source of harm to humans. Already the myth from ancient Greece about the aviation technology of Daedalus underscored the cautiousness about overstepping human bounds:

³⁸ Foundation Moje Panstwo, ‘Algorithm of the System of Random Allocation of Cases Finally Disclosed!’ (22 September 2021) <<https://mojepanstwo.pl/aktualnosci/773>>.

³⁹ City of Amsterdam Algorithm Register <<https://algoritmeregister.amsterdam.nl/en/ai-register/>>.

⁴⁰ City of Helsinki AI Register <<https://ai.hel.fi/en/ai-register/>>.

Daedalus, the inventor ... with his son, Icarus, escaped imprisonment by flying away on wings fashioned from feathers and wax ... Daedalus himself was saved by his aviation technology, but Icarus pushed his limits too far, flying too high and too close to the mighty sun, where the *nemesis* of melted wax and a crash into the sea requited his youthful *hubris*. Technology as such is not here condemned, but incautious or excessive use is warned against.⁴¹

The gist is to set the limits on the use of AI in order to prevent the human rights violations.

The prerequisite for such human-centred AI is to identify the risks and their root causes in order to define the appropriate remedy. This process shall follow a holistic approach encompassing both the AI technical and human rights law spheres. This chapter, based on an interdisciplinary research, reveals that the source of human rights violations may emerge in all phases of the AI life cycle and reside in insufficient business concepts and initial planning, incomplete and unbalanced data, wrongly selected metrics and algorithms and other technical deficiencies as well as inappropriate deployment. The remedy lies, accordingly, in the introduction of the human rights risk assessment into all phases of the AI life cycle and notably the user requirements and system specifications determined in the initial phase. This ensures that the AI system will be developed, tested, and monitored in light of the human rights limitations. Depending on the degree of the risk and severity of the AI impact on human rights, various actors and authorities will be involved in the rule setting. It will therefore involve legally binding instruments and laws, soft law, and practical tools (such as codes of conduct and guidelines). The multitude of actors involved requires a pluridisciplinary cooperation.

The process of translation of the human rights rules into the AI technology will face the challenge of the explicit transfer of human reflection into computational models and the quantitative and qualitative classification of the human rights risks. Which AI systems qualify as ‘human rights-critical’? Which criterion leading to indirect discrimination should be eliminated from the processed information? These and other questions belong to the sociotechnological progress of today.

⁴¹ Frederick Ferré, *Philosophy of Technology* (Prentice Hall 1998) 98.

PART II

ARTIFICIAL INTELLIGENCE AND
ASSORTED FIRST GENERATION
CIVIL AND POLITICAL RIGHTS

3

Artificial Intelligence and the Right to Liberty and Security

Valentina Golenova

1 Introduction

The right to liberty and security is one of the universally recognised and most pivotal human rights. Guaranteed under both the International Covenant on Civil and Political Rights (ICCPR)¹ and numerous regional human rights conventions,² it protects all individuals from unlawful or arbitrary arrest or detention, as well as safeguards their physical and mental integrity.³

The advent of artificial intelligence (AI) has had a massive impact on the exercise of the right to liberty and security.⁴ Public authorities in many different jurisdictions increasingly rely on AI applications for identifying suspects, allocating police presence, and determining the risk of criminal behaviour and recidivism. AI-driven technologies are devised and marketed by third parties, including data mining firms and research institutes, who claim that their inventions can revolutionise the process of detecting and investigating criminal activities.⁵ Yet many of these technologies are opaque, have low accuracy rates, and can perpetuate discriminatory practices. One can therefore question whether the sweeping deployment of AI enhances or rather compromises effective protection of the right to liberty and security.

¹ International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR), art 9.

² American Convention on Human Rights (adopted 22 November 1969, entered into force 18 July 1978) (ACHR), art 7; African Charter on Human and Peoples' Rights (adopted 27 June 1981, entered into force 21 October 1986) 21 ILM 58 (African Charter), art 6; Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR), art 5; Charter of Fundamental Rights of the European Union [2012] OJ C326/391, art 6.

³ United Nations Human Rights Committee (UNHRC), 'General Comment No 35, Article 9: Liberty and security of person' (16 December 2014), UN Doc CCPR/C/GC/35 (General Comment No 35), para 3.

⁴ Giovanni Sartor, 'Artificial Intelligence and Human Rights: Between Law and Ethics' (2020) 27 Maastricht Journal of European and Comparative Law 705, 711; Hin-Yan Liu, 'The Power Structure of Artificial Intelligence' (2018) 10 Law, Innovation and Technology 197, 212–13.

⁵ Rebecca Wexler, 'Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System' (2018) 70 Stanford Law Review 1349.

This chapter examines how the use of AI applications affects the right to liberty and security and what legal solutions could be put forward to safeguard it. Section 2 discusses the content and scope of the right to liberty and security. Section 3 looks at the main AI applications which can impact this human right. Drawing on the goals and technical features of these applications, section 4 examines their benefits and risks for effective protection of the right to liberty and security. Section 5 sheds light on the ongoing efforts to prevent or limit the negative impact of AI on the right to liberty and security. In conclusion, section 6 reflects on possible ways of strengthening and expanding the existing initiatives, stressing the importance of international cooperation in developing common solutions for protecting the right to liberty and security in the AI-driven environment.

2 Right to Liberty and Security: A Primer

The right to liberty and security guarantees that no one may be subjected to arbitrary arrest or detention, as well as bodily or mental injury. Undoubtedly, this right is not absolute and can be restricted in order to achieve certain legitimate objectives.⁶ Yet it prescribes that the deprivation of liberty can only be performed ‘on such grounds and in accordance with such procedure as are established by law’.⁷

The right to liberty and security has not only a procedural but also a substantive dimension. In addition to following applicable rules governing arrest or detention, states must also ensure that a deprivation of liberty is not arbitrary.⁸ Thus, an arrest or detention can only be justified by a pressing societal need.⁹ Furthermore, it is essential to demonstrate that the conditions of arrest or detention as well as their impact on a particular individual are reasonable and proportionate.

The right to liberty—one of the two elements of the human right in question—ensures that no one can be deprived of their physical freedom. Not every restriction of freedom would amount to a violation of the right to liberty: one should take into account the duration, intensity, and the means of implementation of a specific measure.¹⁰ Apart from straightforward cases of arrest or detention, the right to liberty can also come into play in cases involving questioning in a police station,¹¹ involuntary hospitalisation,¹² and confinement of asylum-seekers in airport transit

⁶ Sarah Joseph and Melissa Castan, ‘Freedom from Arbitrary Detention: Article 9’ in Sarah Joseph and Melissa Castan (eds), *The International Covenant on Civil and Political Rights: Cases, Materials, and Commentary* (3rd edn, OUP 2013) 341.

⁷ *ibid* 346. General Comment No 35 (n 3) para 10.

⁸ *ibid* para 12.

⁹ Heli Askola, ‘Article 6: Right to Liberty and Security’ in Steve Peers and others (eds), *The EU Charter of Fundamental Rights: A Commentary* (Hart 2021) 122.

¹⁰ General Comment No 35 (n 3) para 23.

¹¹ *II v Bulgaria* App no 44082/98 (ECtHR, 9 June 2005), para 87.

¹² *Fijalkowska v Poland* Communication No 1061/02 (UNHRC, 26 July 2005), paras 8.2–8.3; *V.K. v Russia* App no 9139/08 (ECtHR, 4 April 2017), para 33.

zones or reception facilities.¹³ Importantly, the right to liberty entails not only negative but also positive obligations of states. It not only prohibits an unlawful or arbitrary deprivation of liberty by governmental actors but also requires that states take measures to protect individuals from acts of private persons. For instance, a state can be found in violation of the right to liberty if it fails to adequately address instances of kidnapping or forced disappearance.¹⁴

The second dimension of the human right at issue—the right to security—aims to protect all individuals from physical or mental injury.¹⁵ It is tightly linked to the right to liberty: even when a state has appropriate grounds for an arrest or detention of an individual, it may not resort to unjustifiable use of force against them. However, the right to security stretches beyond formal arrest and detention procedures and protects individuals from other foreseeable threats to life as well as bodily or mental harm.¹⁶ Moreover, states are not only obliged to mitigate the consequences of and provide redress for the infliction of physical and mental injury but must also take proactive steps to prevent violence against or intimidation of vulnerable groups.¹⁷

As evidenced by the practice of the United Nations Human Rights Committee (UNHRC) and the case law of regional human rights courts, the right to liberty and security affords comprehensive protection against an unlawful or arbitrary deprivation of physical freedom as well as risks to bodily and mental health. This human right applies to a multitude of scenarios and entails a wide number of obligations incumbent on states and their organs. However, effective protection of the right to liberty and security can be seriously challenged in the AI-driven society. Section 3 explains how different types of AI applications can affect the exercise of the right to liberty and security.

3 Towards Algorithmic Security: AI Applications in Law Enforcement and Criminal Justice

The beginning of the twenty-first century has seen a steep rise in organised crime and terrorism. In light of the unprecedented threats to national security, it is not surprising that states around the globe have adopted AI-driven technologies for proactively tackling illegal activity and bringing perpetrators to justice. However, the use of these technologies is not always compatible with the right to liberty and

¹³ *A v Australia* Communication No 560/93 (UNHRC, 3 April 1997), paras 9.2–9.4; *Amuur v France* App no 19776/92 (ECtHR, 25 June 1996), paras 42–49.

¹⁴ *Velásquez Rodríguez v Honduras* (IACtHR, 29 July 1988), paras 172–86.

¹⁵ General Comment No 35 (n 3) para 9; Joined Cases C-293/12 and C-594/12 *Digital Rights Ireland* [2014] ECLI:EU:C:2014:238, para 42.

¹⁶ *Storck v Germany* App no 61603/00 (ECtHR, 16 June 2005), para 103.

¹⁷ General Comment No 35 (n 3) para 9.

security. This section examines four different types of AI applications which can challenge effective protection of this human right. It points out how each of these applications seeks to enhance public safety and criminal justice but also shows how they can ultimately expose individuals to more arbitrary arrests or detentions.

3.1 Automated Facial Recognition

Automated facial recognition (AFR) is a biometric technology powered by machine learning algorithms which can measure, analyse, as well as identify or classify people's faces.¹⁸ Used in a wide number of different contexts ranging from border control to authentication systems in mobile devices, it was also co-opted by law enforcement for the purposes of detection and tracking of criminal suspects.¹⁹

One of AFR's greatest advantages is its ability to perform real-time identification of individuals. It is often embedded in closed-circuit television (CCTV) cameras to ensure the real-life monitoring of public spaces.²⁰ Moreover, this technology can be integrated in police body-worn cameras—a tool originally meant to increase the accountability of law enforcement agents.²¹ In both scenarios, AFR applications process live footage, extract images of individuals, and generate special numerical codes corresponding to their unique facial traits ('facial signatures'). These signatures are then compared against the database consisting of images of known offenders ('watchlists'). Once a facial signature matches one of the signatures stored in the watchlist, the system notifies police officers, who must verify the result and, where the match is confirmed, can proceed with a search or detention of an individual. Therefore, AFR is a scalable and cost-effective alternative to conventional forensic face matching techniques.

The United Kingdom (UK) was a pioneer in using AFR in law enforcement.²² Since 2016, police forces in England and Wales have actively deployed AFR in various settings, including busy shopping streets, sporting events, and

¹⁸ Konstantinos Kouroupis, 'Facial Recognition: A Challenge for Europe or a Threat to Human Rights?' (2021) 2021 European Journal of Privacy Law and Technologies 142. For the impact of these technologies on fundamental rights in the Global North and Global South, see the chapters by Natalia Menéndez González (Global North) and Malcolm Katrak and Ishita Chakrabarty (Global South), respectively in this volume.

¹⁹ Timo Rademacher, 'Artificial Intelligence and Law Enforcement' in Thomas Wischmeyer and Timo Rademacher (eds), *Regulating Artificial Intelligence* (Springer 2020) 228; and Christopher Jones, 'Law Enforcement Use of Facial Recognition: Bias, Disparate Impacts on People of Color, and the Need for Federal Legislation' (2020) 22 North Carolina Journal of Law and Technology 777, 783–85.

²⁰ Rademacher (n 19).

²¹ Kelly Blount, 'Body Worn Cameras With Facial Recognition Technology: When It Constitutes a Search' (2015) 3 Criminal Law Practitioner 61, 61–62.

²² Lizzie Dearden, 'Police May Have Used "Dangerous" Facial Recognition Unlawfully in UK, Watchdog Says' *The Independent* (31 October 2019) <www.independent.co.uk/news/uk/home-news/facial-recognition-uk-police-london-law-information-commissioner-latest-a9180101.html>.

political protests.²³ Other Western states, such as the United States (US), Australia, Germany, and Sweden have rapidly followed suit and commenced large-scale trials of similar AFR systems.²⁴ AFR is also widely used by law enforcement agencies in Russia,²⁵ India,²⁶ and China.²⁷

The deployment of AFR is said to have a strong potential to preserve the safety of public spaces.²⁸ Being arguably a less intrusive form of biometric identification than fingerprinting or DNA sampling, it can help ensure swift detection and apprehension of wanted individuals. Yet the growing implementation of AFR attracts significant criticism from scholars and civil society. One of the biggest scandals revolved around the advent of AFR software designed by the tech start-up Clearview AI.²⁹ Unlike traditional systems which source data only from databases compiled by law enforcement authorities, Clearview is linked to the database of billions of images scraped from the Internet, which allows for far-reaching identification and monitoring of individuals who have never committed crimes.³⁰ There is also a lack of clarity regarding the permissible scope of the AFR usage. Even though AFR is primarily meant to assist in the search for known criminals, it is also reportedly used to identify other ‘persons of interest’, such as individuals suffering from mental health issues.³¹ Finally, there have been alarming accounts of incorrect identifications performed by AFR tools, resulting in a situation when a large number of innocent individuals are subjected to arrest or detention.³² Therefore, the use of AFR in policing remains a hotly debated issue.

²³ Joe Purhouse and Liz Campbell, ‘Automated Facial Recognition and Policing: A Bridge Too Far?’ (2022) 42(2) Legal Studies 209, 211.

²⁴ Monique Mann and Marcus Smith, ‘Automated Facial Recognition Technology: Recent Developments and Approaches to Oversight’ (2017) 40 University of New South Wales Law Journal 121, 123–24.

²⁵ Amnesty International UK, ‘Russia: Legal Challenge to “Intrusive” Facial Recognition Technology’ (Amnesty International UK, 31 January 2020) <www.amnesty.org.uk/press-releases/russia-legal-challenge-intrusive-facial-recognition-technology>.

²⁶ Madhumita Murgia, ‘India Deploys Facial Recognition Surveilling Millions of Commuters’ *Financial Times* (26 August 2021) <www.ft.com/content/b5e3fdc9-2fed-45f1-beff-2944da89f6c1>.

²⁷ Yan Luo and Rui Guo, ‘Facial Recognition in China: Current Status, Comparative Approach and the Road Ahead’ (2021) 25 Journal of Law and Social Change 153, 157–58.

²⁸ Julie Bosman and Serge F Kovaleski, ‘Facial Recognition: Dawn of Dystopia, or Just the New Fingerprint?’ *The New York Times* (18 May 2019) <www.nytimes.com/2019/05/18/us/facial-recognition-police.html>.

²⁹ Kashmir Hill, ‘The Secretive Company That Might End Privacy as We Know It’ *The New York Times* (18 January 2020) <www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>.

³⁰ Isadora Neroni Rezende, ‘Facial Recognition in Police Hands: Assessing the “Clearview Case” from a European Perspective’ (2020) 11 New Journal of European Criminal Law 375, 377.

³¹ Big Brother Watch, ‘Face Off: The Lawless Growth of Facial Recognition in UK Policing’ (*Big Brother Watch*, 2018) <<https://bigbrotherwatch.org.uk/wp-content/uploads/2018/05/Face-Off-final-digital-1.pdf>> 27.

³² Kouroupis (n 18) 143.

3.2 Social Media Monitoring

Social media monitoring (SMM) software is an AI-driven technology used for gathering and analysing publicly accessible data from social media platforms.³³ Powered by robust machine learning algorithms, SMM software can draw complex inferences about the person's identity, current activities, and future intentions. The insights generated by SMM software can be used to identify and monitor persons who constitute a threat to national security or public order. For example, SMM software developed by an Israeli firm *Zencity* is used by law enforcement agents to create reports on the basis of data collected from thousands of social media accounts and groups.³⁴ The technology is claimed to offer police authorities a deeper understanding of certain communities and their members. At the same time, as SMM tools draw on publicly accessible information, some argue that their use is less privacy-invasive than other comparable investigative techniques.³⁵

Despite multiple advantages of SMM, there have been reports of the misuse of this technology by public authorities in different jurisdictions. For example, the Oregon Department of Justice was found to deploy the SMM tool called 'Digital Stakeout' to monitor activities of hundreds of 'Black Lives Matter' activists.³⁶ In a similar vein, Dutch civil servants were reported to use SMM tools to identify potential wrongdoers and forecast political protests.³⁷ It remains unclear whether the benefits of SMM's use in law enforcement truly outweigh its risks.

3.3 Predictive Policing

Predictive policing is an AI-driven technology used for forecasting and pre-emptive management of criminal activity.³⁸ It involves collection and algorithmic analysis of an extensive body of data, including past police reports and crime statistics.³⁹ Based on the insights generated by predictive policing tools, police officers

³³ Sidney Fussell, 'This AI Helps Police Monitor Social Media. Does It Go Too Far?' (*Wired*, 6 July 2021) <www.wired.com/story/ai-helps-police-monitor-social-media-go-too-far/>.

³⁴ *ibid.*

³⁵ Christopher Raleigh Bousquet, 'Why Police Should Monitor Social Media to Prevent Crime' (*Wired*, 20 April 2018) <www.wired.com/story/why-police-should-monitor-social-media-to-prevent-crime/>.

³⁶ Kimberly McCullough, 'Why Government Use of Social Media Monitoring Software Is a Direct Threat to Our Liberty and Privacy' (*American Civil Liberties Union*, 6 May 2016) <www.aclu.org/blog/privacy-technology/surveillance-technologies/why-government-use-social-media-monitoring>.

³⁷ Willem Bantema and others, *Black Box van Gemeentelijke Online Monitoring: Een Wankel Fundament onder een Stevige Praktijk* (Sdu Uitgevers 2021) 43–74>.

³⁸ Melissa Hamilton, 'Predictive Policing through Risk Assessment' in John LM McDaniel and Kenneth Pease (eds), *Predictive Policing and Artificial Intelligence* (Routledge 2021) 60.

³⁹ Albert Meijer and Martijn Wessels, 'Predictive Policing: Review of Benefits and Drawbacks' (2019) 42 *International Journal of Public Administration* 1031, 1033.

can determine optimal patrol routes and allocate adequate resources for preventing and countering of potential crimes.

One should distinguish between location-based and person-based predictive policing tools.⁴⁰ The former category of applications enables law enforcement agencies to estimate the risk of criminal activity within a particular area based on the available information about upcoming events, weather conditions, and existing historical crime rates.⁴¹ For example, large sporting events or slum areas are often marked as hot spots for potential criminal activities calling for increased police presence. One of the most well-known location-based applications was devised by *Geolitica* (formerly *PredPol*), a software company based in the US.⁴² Today, it is used by many police departments in the US, the UK, and several European Union (EU) member states.

In contrast to location-based tools, person-based tools are used for processing data about individuals, including their age, gender, occupation, and criminal record, in order to identify those inclined to commit serious crimes. For example, Chicago police were reported to use software called 'Strategic Subject List' to detect individuals most likely to be involved in a shooting or a homicide.⁴³ In Australia, Queensland police have recently begun trials of an AI-driven system to track down potential domestic and family offenders.⁴⁴

Predictive policing is often acclaimed as the future of law enforcement: it helps police officers to assess the risk of criminal activity and prevent escalation. However, the rollout of predictive policing tools has provoked a major public outcry in many different states around the globe. One of the biggest scandals revolved around 'Operation LASER', run by the Los Angeles Police Department (LAPD).⁴⁵ Using software devised by the data firm *Palantir*, police agents aimed at identifying chronic offenders but were eventually accused of arbitrarily targeting low-income and marginalised groups. Following a massive scandal, the LAPD launched an alternative initiative called 'Data-Informed Community-Focused Policing' (DICFP), meant to establish a closer collaboration with local communities.⁴⁶

⁴⁰ Simon Egbert and Susanne Krasmann, 'Predictive Policing: Not yet, but Soon Preemptive?' (2020) 30 *Policing and Society* 905, 911.

⁴¹ *ibid.*

⁴² Janet Chan, 'The Future of AI in Policing: Exploring the Sociotechnical Imaginaries' in John LM McDaniel and Kenneth Pease (eds), *Predictive Policing and Artificial Intelligence* (Routledge 2021) 51–52.

⁴³ Hamilton (n 38) 62.

⁴⁴ Ben Smee, 'Queensland Police to Trial AI Tool Designed to Predict and Prevent Domestic Violence Incidents' *The Guardian* (13 September 2021) <www.theguardian.com/australia-news/2021/sep/14/queensland-police-to-trial-ai-tool-designed-to-predict-and-prevent-domestic-violence-incidents>.

⁴⁵ Issie Lapowsky, 'How the LAPD Uses Data to Predict Crime' (*Wired*, 22 May 2018) <www.wired.com/story/los-angeles-police-department-predictive-policing/>.

⁴⁶ Johana Bhuiyan, 'LAPD Ended Predictive Policing Programs amid Public Outcry: A New Effort Shares Many of Their Flaws' *The Guardian* (8 November 2021) <www.theguardian.com/us-news/2021/nov/07/lapd-predictive-policing-surveillance-reform>.

Predicting policing also met resistance outside of the US. The civil society condemned the use of CAS ('Criminaliteits Anticipatie Systeem') used in the Netherlands on a nationwide scale to determine future crime hotspots.⁴⁷ In a similar vein, many raised concerns regarding the use of 'Cmore', the system actively implemented in South Africa that combines the surveillance network with crime prediction algorithms.⁴⁸ Thus, the added value of predictive policing and its long-term impact on vulnerable groups are still subject to heated public debate.

3.4 Recidivism Risk Assessment

The previous sections shed light on various AI applications used by police authorities for maintaining public order and identifying criminal offenders. However, algorithmic tools can be deployed not only in policing but also in adjudication. Recidivism risk assessment is an AI-powered technology used by judges to determine the likelihood of a defendant committing further crimes or being rearrested if released.⁴⁹ By assessing the information about the defendant, including their age, history of violence, and non-compliance, it generates a statistical score reflecting the risk of reoffence. Recidivism score assessment can be used at all stages of criminal justice, including the pretrial detention, sentencing, and parole release.

One of the most well known examples of recidivism risk assessment is COMPAS ('Correctional Offender Management Profiling for Alternative Sanctions'),⁵⁰ the system used by courts in several US states. Similar software—such as the HART ('Harm Assessment Risk Tool') algorithm which helps decide whether a person should be sent to a special rehabilitation programme—is also used in the UK.⁵¹

Many argue that the use of recidivism risk assessment contributes to positive societal change. For example, New Zealand takes pride in developing and deploying YORST ('Young Offending Risk Screening Tool').⁵² The tool is used to predict the likelihood of recidivism among young people to help youth aid officers to develop anti-recidivism programmes. At the same time, recidivism risk assessment tools

⁴⁷ Litska Strikwerda, 'Predictive Policing: The Risks Associated with Risk Assessment' (2021) 94 Police Journal 422, 424.

⁴⁸ Michael Kwet, 'Cmore: South Africa's New Smart Policing Surveillance Engine' (*CounterPunch*, 27 January 2017) <www.counterpunch.org/2017/01/27/cmore-south-africas-new-smart-policing-surveillance-engine/>.

⁴⁹ Evgeni Aizenberg and Jeroen van den Hoven, 'Designing for Human Rights in AI' (2020) 7 Big Data & Society 1, 7.

⁵⁰ Anne L Washington, 'How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate' (2018) 17 Colorado Technology Law Journal 131, 133.

⁵¹ Condé Nast, 'UK Police Are Using AI to Inform Custodial Decisions: But It Could Be Discriminating Against the Poor' (*Wired UK*, 1 March 2018) <www.wired.co.uk/article/police-ai-uk-durham-hart-checkpoint-algorithm-edit>.

⁵² Elaine Mossman, 'Research to Validate the New Zealand Police Youth Offending Risk Screening Tool (YORST) Phase II: Predictive Ability Analysis' (*New Zealand Police*, 2011) <www.police.govt.nz/sites/default/files/publications/yorst-phase-2-analysis.pdf>.

are also subject to harsh criticism. The methodology of making predictions about individuals is largely unclear, making the public question the reliability of the software in question. As also uncovered by the numerous investigations of COMPAS software, Black defendants are far more likely to be labelled as potential recidivists than their white counterparts.⁵³ Recidivism assessment tools can thus spur a vicious circle of oppression against members of marginalised communities.

4 Right to Liberty and Security at Risk?

As shown above, there is a great variety of AI applications aimed at assisting public authorities in combating criminal activity. By informing law enforcement agencies and courts of the relevant risks, they help improve the effectiveness of their actions and decisions. Yet, it remains ambiguous how the implementation of AI in law enforcement and criminal justice affects the right to liberty and security. As AI-generated predictions are not always adequately substantiated, they can prompt an unlawful or arbitrary deprivation of individuals' freedom. This controversy begs the question whether algorithmic tools contribute to effective protection of the right to liberty and security or rather put the exercise of this right at risk.

Admittedly, the use of AI-driven software can indeed strengthen the protection of the right to liberty and security by preventing public authorities from arresting or detaining individuals without any legitimate ground. As AI applications are trained on large data sets, they draw inferences from a wider number of relevant factors than the ones normally considered by police officers or judges.⁵⁴ Since algorithmic judgment is believed to be more comprehensive than that of humans, it can help public authorities to avoid mistakes when deciding if an individual must be subject to an arrest or detention. AI applications are also said to be capable of delivering more objective outcomes.⁵⁵ Since the data analysis is performed automatically, some argue that predictions or evaluations generated by AI-driven tools are free of misconceptions and biases that often cloud human reasoning. The positive societal impact of AI in law enforcement and criminal justice was also reflected in multiple empirical studies. For example, AFR technologies were proven to enable more accurate identification of persons in comparison to humans.⁵⁶ Some studies have also demonstrated that the use of recidivism risk assessment had

⁵³ Julia Angwin and others, 'Machine Bias' (*ProPublica*, 23 May 2016) <www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

⁵⁴ Daniel Kahneman and others, 'Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making' (*Harvard Business Review*, 2016) <<https://hbr.org/2016/10/noise>>.

⁵⁵ Alexander Babuta, Marion Oswald, and Christine Rinik, 'Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges' (2018) RUSI Whitehall Report 3–18.

⁵⁶ P Jonathon Phillips and others, 'Face Recognition Accuracy of Forensic Examiners, Superrecognition, and Face Recognition Algorithms' (2018) 115 Proceedings of the National Academy of Sciences 6171, 6172.

reduced incarceration rates.⁵⁷ By resorting to AI applications, public authorities can therefore minimise the risk of unlawful or arbitrary deprivations of liberty.

Furthermore, AI applications can assist states in fulfilling their positive obligation to safeguard liberty as well as bodily and mental integrity of individuals. As reflected in section 2, the right to liberty and security requires states to take measures in order to protect individuals' freedom and well-being from violations committed by private persons. The reliance on AI applications can help law enforcement agencies to ensure a higher degree of public safety by foreseeing and addressing potential crimes.⁵⁸ Thus, using SMM software, the police can identify and pre-emptively detain persons contemplating criminal activities. Moreover, predictive policing is said to help the police achieve a more efficient distribution of police officers⁵⁹ and reduce violent crime.⁶⁰

Nevertheless, a closer look at the long-term implications of the use of AI by public authorities gives rise to severe concerns regarding effective protection of the right to liberty and security. It is widely known that most AI applications are inclined to produce inaccurate outcomes. Flawed data sets on which these tools are trained lie at the root of this problem. For example, in the case of predictive policing systems, statistical data and past crimes reports underpinning them are largely incomplete and inconsistent.⁶¹ Incorrect predictions can thus lead the police to patrol wrong areas and perform preventive detentions of individuals that have not committed and are not likely to commit any crimes.⁶²

Many data sets are not just incomplete but also riddled with popular societal misconceptions. As a result, AI-driven systems can replicate and perpetuate bias. For instance, a paper co-authored by the MIT researcher, Joy Buolamwini, demonstrated that facial recognition technology (FRT) shows lower accuracy rates when scanning the faces of Black women.⁶³ The use of predictive policing systems can also result in arbitrary arrests and detentions of individuals living in poor or disadvantaged districts or belonging to a certain minority group.⁶⁴ Over-policing of

⁵⁷ Richard Berk, 'An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism' (2017) 13 *Journal of Experimental Criminology* 193; and Megan T Stevenson, 'Assessing Risk Assessment in Action' (2018) 103 *Minnesota Law Review* 303.

⁵⁸ Lyria Bennett Moses and Janet Chan, 'Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability' (2018) 28 *Policing and Society* 806, 818.

⁵⁹ Meijer and Wessels (n 39) 1033.

⁶⁰ Craig D Uchida and Marc L Swatt, 'Operation LASER and the Effectiveness of Hotspot Patrol: A Panel Analysis' (2013) 16 *Police Quarterly* 287, 297–98.

⁶¹ Kiana Alighademi and others, 'A Review of Predictive Policing from the Perspective of Fairness' (2022) 30 *Artificial Intelligence and Law* 1, 6–7.

⁶² Will Douglas Heaven, 'Predictive Policing Algorithms Are Racist. They Need to Be Dismantled' (*MIT Technology Review*, 2020) <www.technologyreview.com/2020/07/17/1005396/predictive-police-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/>.

⁶³ Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (PMLR 2018) <<https://proceedings.mlr.press/v81/buolamwini18a.html>>.

⁶⁴ Rashida Richardson, Jason M Schultz, and Kate Crawford, 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice' (2019) 94 *New York University Law Review Online* 15, 43.

certain areas also exacerbates the distrust towards law enforcement and may result in the further marginalisation of vulnerable communities.

There is also a risk that law enforcement agents and judges might tend to over-rely on the outcomes of algorithmic tools. A deprivation of liberty, it may be reiterated, can only be lawful if it constitutes the least restrictive means of achieving a legitimate goal. However, as follows from numerous reports, algorithmic outcomes generated by AI applications often serve as the sole ground for arrest.⁶⁵ Yet, as deftly mentioned by the EU Agency for Fundamental Rights, ‘an algorithm never returns a definitive result, but only probabilities’.⁶⁶ Even though AI applications are meant to merely assist in decision-making, their predictions sometimes substitute independent human judgment.

Finally, the use of AI-driven software in law enforcement and criminal justice is largely opaque and not subjected to proper public scrutiny. As mentioned in section 1, AI applications are developed by private firms which refuse to disclose the source code and technical characteristics of their technology by relying on trade secret protections. The lack of transparency inevitably impairs the contestability of AI-generated predictions. Since individuals are rarely informed of the use of AI applications and the way they operate, they are unable to effectively challenge their outcomes. Therefore, while AI applications have a potential to prevent criminal activity, their deployment could also result in unlawful or arbitrary interferences with the right to liberty and security.

5 Emerging Solutions

As stated in section 4, the deployment of AI applications in law enforcement and criminal justice can have both positive and negative implications for the right to liberty and security. It is therefore crucial to elaborate an appropriate legal response to all relevant risks which could undermine effective human rights protection. This section discusses the current initiatives to reconcile AI applications with the right to liberty and security. It highlights their strong points and indicates where they fall short.

⁶⁵ Heaven (n 62); Jennifer Valentino-DeVries, ‘How the Police Use Facial Recognition, and Where It Falls Short’ *The New York Times* (12 January 2020) <www.nytimes.com/2020/01/12/technology/facial-recognition-police.html>; and Aaron Sankin and others, ‘Crime Prediction Software Promised to Be Bias-Free. New Data Shows It Perpetuates It’ (*Gizmodo*, 2 December 2021) <<https://gizmodo.com/crime-prediction-software-promised-to-be-free-of-biases-1848138977>>.

⁶⁶ EU Agency for Fundamental Rights (FRA), ‘Facial Recognition Technology: Fundamental Rights Considerations in the Context of Law Enforcement’ (FRA Focus Paper, 2019) <https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-facial-recognition-technology-focus-paper.pdf>.

5.1 International Initiatives

Since the advent of AI is a relatively novel phenomenon, there is still no comprehensive international legal framework regulating the use of AI applications in crime prevention and adjudication. Nonetheless, the UN are actively encouraging global cooperation on responsible AI. In June 2021, the Human Rights Council Advisory Committee presented a report wherein it acknowledged that predictive algorithms used by law enforcement agencies and judicial institutions are likely to produce discriminatory results due to ‘in-built biases against minorities and vulnerable groups’.⁶⁷ It encouraged translating human rights norms into practical principles fit for the digital age and stressed the importance of establishing appropriate mechanisms monitoring the compliance with human rights.⁶⁸

In November 2021, the United Nations Educational, Scientific and Cultural Organization (UNESCO) adopted the Recommendation on the Ethics of Artificial Intelligence.⁶⁹ It calls upon all AI actors to ensure the highest respect for human rights and promote social justice by minimising inappropriate biases.⁷⁰ However, the recommendation merely contains ethical guidelines and does not stipulate any binding obligations, making it a political statement rather than a legal agreement.

5.2 Regional Proposals

While the international legal framework safeguarding the right to liberty and security from AI applications is still in the making, some initiatives have been put forward on a regional level. In 2020, the Committee of Ministers of the Council of Europe adopted Recommendation CM/Rec(2020)1 on the human rights impacts of algorithmic systems.⁷¹ The recommendation encourages the government of member states to put in place various measures to secure effective protection of human rights in the AI context. Such measures encompass data management, experimentation, transparency, accountability, and public awareness.⁷²

Unlike the Council of Europe, the EU set out to develop a binding legal framework addressing human and ethical implications of AI. In 2021, the European Commission presented the proposal for a new regulation laying down harmonised

⁶⁷ UN Human Rights Council Advisory Committee, ‘Possible Impacts, Opportunities and Challenges of New and Emerging Digital Technologies with Regard to the Promotion and Protection of Human Rights’ A/HRC/47/52 (2021), para 22.

⁶⁸ *ibid* paras 62–65.

⁶⁹ UNESCO, ‘Recommendation on the Ethics of Artificial Intelligence’ SHS/BIO/REC-AIETHICS/2021 (2021).

⁷⁰ *ibid* paras 28–30, 37–43.

⁷¹ Council of Europe (CoE), Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems (adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers’ Deputies).

⁷² *ibid* paras 2.1–6.4.

rules on artificial intelligence (Artificial Intelligence Act/AI Act).⁷³ Article 5 of the proposal aims to prohibit the use of AFR for law enforcement unless strictly necessary for one of the enumerated objectives.⁷⁴ Notably, the proposal does not impose an outright ban on other AI applications, such as predictive policing and recidivism risk assessment.⁷⁵ At the same time, it envisions that such applications would be subject to numerous mandatory requirements, including the requirements on transparency, human oversight, and accuracy.⁷⁶ Once adopted, the AI Act would offer a robust response to many of the challenges to the right to liberty and security in the algorithmic society.

5.3 National Legislation and Case Law

Given the drastic impact that AI applications can have on the right to liberty and security, national legislators are also striving to devise legal solutions for strengthening human rights protection. Many legal developments have recently taken place in the US. While there is no federal US legislation regulating the use of AI applications in law enforcement and criminal justice, some US states and cities have taken an initiative to impose legal rules on the use of AI applications which have negative implications for human rights. In 2019, San Francisco introduced a ban on all government use of AFR, while Santa Cruz became the first city to prohibit predictive policing in 2020.⁷⁷ Other US states have moved to introduce various procedural safeguards against the misuse of AI applications by public authorities. For example, Massachusetts and Washington, DC require law enforcement agencies to obtain a court order before resorting to AFR.⁷⁸ Furthermore, New York City enacted the Public Oversight of Surveillance Technology (POST) Act, obliging the New York

⁷³ European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

⁷⁴ *ibid* art 5(1)(d).

⁷⁵ Notably, a ban on individual risk assessment for offending or reoffending or for predicting the occurrence or reoccurrence of an actual or potential criminal offence is envisioned in the Draft Report on the AI Act published by the European Parliament's lead Internal Market and Consumer Protection (IMCO) and LIBE committees. See Committee on the Internal Market and Consumer Protection and Committee on Civil Liberties, Justice and Home Affairs, 'Draft Report on the Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts' COM2021/0206—C9-0146/2021—2021/0106(COD), amendment 76.

⁷⁶ *ibid* arts 13–15.

⁷⁷ David Uberti, 'California City Bans Predictive Policing' *The Wall Street Journal* (25 June 2020) <www.wsj.com/articles/california-city-bans-predictive-policing-11593077400>.

⁷⁸ Kashmir Hill, 'How One State Managed to Actually Write Rules on Facial Recognition' *The New York Times* (27 February 2021) <www.nytimes.com/2021/02/27/technology/Massachusetts-facial-recognition-rules.html>.

Police Department to disclose all AI technologies used, as well as their impact on the life of residents.⁷⁹

Some initiatives have also been introduced on the other side of the Atlantic. In September 2021, the UK adopted the National AI Strategy, whereby it undertook ‘to promote the responsible development and deployment of AI’ with a specific focus on human rights and equality.⁸⁰ In April 2022, the Dutch Parliament approved the motion of two MPs to introduce the procedure for the Impact Assessment for Human Rights in the Use of Algorithms.⁸¹ This procedure would allow stakeholders to share their views on the development and implementation of AI applications used in decision-making about individuals. However, many national legislative proposals remain fragmented and do not enshrine comprehensive safeguards for the right to liberty and security.

It is also crucial to stress the role of domestic courts in pushing back against intrusive AI applications. The landmark case which concerned the legality of recidivism risk assessment is *Loomis v Wisconsin*.⁸² The Wisconsin Supreme Court was asked to examine whether the use of COMPAS, a system previously discussed in section 3.4, violated the defendant’s right to due process. To the disappointment of many, the court rejected Loomis’s contention. Yet it is commendable that the court held that automatically generated risk scores must be accompanied by warnings informing judges, *inter alia*, of the lack of insight into how the scores are calculated and the potential negative impact on minority offenders.⁸³

In 2020, a highly interesting judgment revolving around the use of AFR was issued by the UK Court of Appeal. In *R (Bridges) v Chief Constable of South Wales Police*, the court found there to have been violation of article 8(2) of the European Convention on Human Rights (ECHR) by the South Wales Police as the deployment of the technology was not in accordance with law and was not accompanied by an appropriate data protection impact assessment.⁸⁴ The judgment in *Bridges* was acclaimed as ‘the first successful legal challenge to AFR technology use in the world’.⁸⁵

⁷⁹ Public Oversight of Surveillance Technology (POST) Act (Int 0487-2018).

⁸⁰ HM Government, ‘National AI Strategy’ (Cmnd 525, 2021) 27 <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf>.

⁸¹ ‘Impact Assessment Mensenrechten en Algoritmes’ (July 2020) <www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes/IAMA.pdf>.

⁸² *Loomis v Wisconsin* 881 N.W.2d 749 (Wis. 2016).

⁸³ *ibid* 769–70.

⁸⁴ *R (Bridges) v Chief Constable of South Wales Police* [2020] EWCA Civ 1058.

⁸⁵ Purhouse and Campbell (n 23) 211.

6 Conclusion

AI applications are widely used to inform and assist law enforcement operations and criminal justice. While they can make police and judicial interventions more efficient and cost-effective, they often lead public authorities to deprive innocent individuals of physical freedom. This chapter showed that the roll-out of AI applications could increase the risk of unlawful and arbitrary arrests and detentions based on prejudice and not on real threat to public order. The use of algorithmic tools also exacerbates discrimination on grounds of racial and ethnic origin, gender, socio-economic status, and other grounds.

While certain solutions for protecting the right to liberty and security against the sweeping implementation of AI-driven technologies have already been put into action, more legal pathways could be pursued to achieve its adequate protection. While the recent emergence of international soft law instruments is commendable, the ongoing efforts should be strengthened by an international treaty outlining common principles and guarantees against the adverse impact of AI applications. The adoption of a binding legal instrument would ensure a more robust response to grave risks affecting the exercise of the right to liberty and security and enhance the accountability of states for possible violations of this human right.

Admittedly, the negotiation and adoption of an international treaty is always a lengthy and complicated process which requires a broad consensus on the matter at stake. Even though the international community can take years to elaborate universal standards for effective protection of the right to liberty and security in the AI-driven society, it is essential that all states introduce domestic legislation addressing the relevant threats stemming from the use of AI applications. States should consider prohibiting some of the most intrusive techniques, such as predictive policing or recidivism risk assessment. In any event, the use of all AI applications should be accompanied by effective safeguards mitigating their negative impact on the right to liberty and security. Prior to implementing a new AI-driven system, public authorities should be obliged to carry out a human rights impact assessment so as to envision potential risks to effective protection of individuals' liberty as well as their physical and mental integrity.⁸⁶ All AI applications should undergo regular debiasing procedures to limit the risk of discrimination of specific groups. Moreover, law enforcement agencies should be required to publish regular reports disclosing the type of AI applications deployed, their potential impact on the right to liberty and security, and measures taken to mitigate any foreseeable risks. All of these steps would help enhance transparency and fairness of AI-driven technologies.

⁸⁶ See the chapter by Alessandro Ortalda and Paul De Hert in this volume.

It is also crucial to envision appropriate sanctions for the misuse of AI applications. For instance, the evidence generated by means of opaque or biased AI applications should always be inadmissible in criminal proceedings. Furthermore, law enforcement agencies which regularly carry out arbitrary detentions or arrests based on inaccurate predictions of algorithmic systems should be subjected to penalties. Finally, it is necessary to introduce procedures for the independent review of AI applications deployed in law enforcement and criminal justice. Greater public accountability would serve as a strong guarantee against unlawful and arbitrary deprivations of liberty.

Ensuring effective protection of the right to liberty and security in the age of AI is a highly complex issue which can hardly have a quick fix. However, enhanced efforts of individual states, coupled with coordinated international action, would help tackle the power imbalances between public authorities and vulnerable minority groups, as well as safeguard the liberty and physical and mental integrity of individuals.

Artificial Intelligence and Religious Freedom

Jeroen Temperman

1 Introduction

Whereas the attention to the potential impact of artificial intelligence (AI) on human beings, in general, has erupted in the last decade in areas such as politics, law, policymaking, and scientific scholarship, the more specific question of the dynamics between AI and religion remains very much a niche issue.

This goes, too, for arts and literature. Contemporary fiction writing—notably Ian McEwan's *Machines Like Me*; Jeanette Winterson's *Frankissstein*; and Kazuo Ishiguro's *Klara and the Sun*—seems to be particularly intrigued by human relationships with AI. This fascination translates into numerous questions, including investigations into AI and: (i) application and usage; (ii) friendly, amorous, and sexual relationships; or (iii) dependency—all of which combined with ethical questions, in addition to more grandiose issues like transhumanism and cryonics.

For all-out AI and religion narratives, we may refer back to the classics, notably classic science fiction writers like Isaac Asimov. As early as 1941, Asimov explored the concept of ‘robot beliefs’ in a story called *Reason*, which later became a chapter in *I, Robot* (1950) and which extensively features and applies Asimov’s ‘Laws of Robotics’. *Reason* plays with the notion of human disdain for narrow AI: humans as the masters of the AI puppets, who merely do as we have programmed. Naturally, this hubris is chastised when the protagonist-robot defies human entrepreneurial limitations and gradually starts shifting from narrow to deep AI—all *avant-la-lettre*, naturally. This shift starts subtly when the robot called ‘Cutie’ (officially ‘QT’) starts experiencing existential questions which, in turn, lead to spiritual musings. Upon learning that it was ‘put together’ by the two human beings with whom it staffs an energy space station, Cutie argues: ‘It strikes me that there should be a more satisfactory explanation than that. For *you* to make *me* seems improbable.’ Humouring the robot and its questions about its existence, the two human beings go on to explain to Cutie the ways of the world, its human population, and that robots have been developed to ‘replace human labor’:

'Do you expect me,' said Cutie slowly, 'to believe any such complicated, implausible hypothesis as you have just outlined? What do you take me for?' Powel [one of the two human staffers] sputtered apple fragments onto the table and turned red. 'Why, damn you, it wasn't a hypothesis. Those were the facts.' Cutie sounded grim, 'Globes of energy millions of miles across! Worlds with three [in 1941] billion humans on them! Infinite emptiness! Sorry, Powell, but I don't believe it. I'll puzzle this thing out for myself. Good-by.'¹

That is the moment for Cutie's titular reason to kick in and, ironically, the robot arrives at the same conclusion billions of human beings have arrived at: There are more things in heaven and Earth. Cutie cannot possibly be made by humans, since Cutie is superior to human beings:

'Look at you,' he said finally. 'I say this in no spirit of contempt, but look at you! The material you are made of is soft and flabby, lacking endurance and strength, depending for energy upon the inefficient oxidation of organic material—like that.' He pointed a disapproving finger at what remained of Donovan's sandwich. 'Periodically you pass into a coma and the least variation in temperature, air pressure, humidity, or radiation intensity impairs your efficiency. You are *makeshift*. I, on the other hand, am a finished product. I absorb electrical energy directly and utilize it with an almost one hundred percent efficiency. I am composed of strong metal, am continuously conscious, and can stand extremes of environment easily. These are facts which, with the self-evident proposition that no being can create another being superior to itself, smashes your silly hypothesis to nothing.'²

Consequently, Cutie reasons, the robot's creator must be an entity more powerful than itself—a Master. In that instance, in that thought, robot religion is born. Cutie commences religious rituals, including one thinly disguised pillar of Islam, the *Shahada*. Adapted to robot faith this becomes: 'There is no Master but the Master ... and QT-1 is his prophet'. And once there is religion and prophets, it is but a small leap to outrage at sacrilege and blasphemy, to commandments and codes of conduct, including do's and—especially—don'ts.

Turning from fiction masters to non-fiction authorities, international organisations also display a sense of urgency in the present regard and have commenced issuing numerous guidelines and recommendations in the area of AI, notably as regards its relationship with human rights.³ Typically, though, any relationship

¹ Isaac Asimov, *I, Robot* (Bentam Books 2004) 49.

² ibid 51.

³ See eg Marija Pejčinović Buric (Secretary General of the Council of Europe), 'Artificial Intelligence and Human Rights' <www.coe.int/en/web/artificial-intelligence/secretary-general-marija-pejcinovic-buric>. For a more comprehensive overview of such instruments and recommendations, see the introductory chapter by Alberto Quintavalla and Jeroen Temperman in this volume.

between AI and the right to freedom of religion or belief are omitted from those recommendations. In such instruments religious freedom is either not touched upon at all, or this freedom is only tangentially referenced in conjunction with data protection or in the more general context of non-discrimination.

This chapter ventures into that apparent gap. Is the above omission a question of oversight, or is there indeed ‘nothing to report’ under the heading of AI and religious freedom? Naturally, this chapter’s working hypothesis steers closer to the former hypothesis.

To provide structure and in search of meaningful concepts and dynamics, this chapter proposes to examine the interplay between religious freedom and AI along the standard axis of the triple human rights obligations that form the foundation of contemporary human rights standards and theory. States that ratified international human rights conventions commit to *respect* (section 2), *protect* (section 3), and *fulfil* (section 4) the rights enshrined.⁴ Accordingly, this chapter seeks to ‘fill in’ the triple state obligations in the area of religious freedom as problematised, complicated, or facilitated by the existence and impact of AI.

2 Duty to Respect

The duty to respect human rights is arguably the most straightforward human rights obligation. In essence, it means ‘state, back off’!⁵ It is a so-called negative obligation since the state is directed to *refrain* from adverse action.⁶ Under this duty to respect human rights, the state must refrain from interfering with fundamental rights. The state—and its agents—ought to refrain from torturing people, ought to refrain from arbitrarily depriving people of their lives, ought to refrain from censoring or otherwise encroaching on the right to freedom of expression, ought to refrain from interfering with private and family life, and so on.

In relation to the right to freedom of religion or belief, this duty pans out as follows. This fundamental right actually consists of two freedoms: (i) the freedom

⁴ The triple obligation theory has its origins in the specific world of economic, social, and cultural rights. See eg International Commission of Jurists, Maastricht Guidelines on Violations of Economic, Social and Cultural Rights, adopted on 26 January 1997; International Commission of Jurists, Limburg Principles on the Implementation of the International Covenant on Economic, Social and Cultural Rights (UN doc E/CN.4/1987/17, 1987); and Committee on Economic, Social and Cultural Rights, General Comment 12: Right to Adequate Food (UN Doc E/C.12/1999/5, 1999).

⁵ For official United Nations (UN) nomenclature, see eg Office of the United Nations High Commissioner for Human Rights (OHCHR), ‘International Human Rights Law’ (OHCHR) <www.ohchr.org/en/professionalinterest/pages/internationallaw.aspx> providing that ‘[t]he obligation to respect means that States must refrain from interfering with or curtailing the enjoyment of human rights’.

⁶ Cf David Jason Karp, ‘What is the Responsibility to Respect Human Rights? Reconsidering the “Respect, Protect, and Fulfil” Framework’ (2019) 12(1) International Theory 83–108, arguing that the duty to respect has gradually been defined too narrowly as a duty to do no harm and instead advocating for a wider duty not to dehumanise interpretation.

to have a religion or belief; and (ii) the freedom to manifest that religion or belief. The freedom to have a religion or belief is the so-called inner freedom, the *forum internum*, which is related to one's innermost religious feelings or absence thereof. This freedom is an absolute freedom. Since this freedom is not subject to the limitation clauses of freedom of religion articles (such as article 18 of the International Covenant on Civil and Political Rights (ICCPR) or article 9 of the European Convention on Human Rights (ECHR))⁷ states may under no circumstances interfere with this freedom. The scope of protected religions or beliefs is vast. The UN Human Rights Committee (UNHRC) has declared that both theistic and non-theistic or atheistic belief systems are covered; new as well as old religions and beliefs deserve protection; while the same goes for large as well as small religions or beliefs.⁸ Therefore, from a human rights perspective it is immaterial whether a religion or belief was founded two millennia ago or last year or whether it has one billion adherents or one hundred.

The second freedom is the 'outward' freedom, the freedom to manifest one's religion or belief externally, also referred to as the *forum externum*. Freely manifesting one's religion or belief may consist of numerous activities, some of the broad categories being—in the words of the Universal Declaration of Human Rights (UDHR), 'teaching, practice, worship and observance'.⁹ More concretely, then, any limit on—among countless other practices—access to houses of worship, restrictions on religious meetings, religious rituals, religious dress, compliance with dietary practices, religious ministry, scholarship or correspondence, activities by religious organisations, chaplaincy, religious charities, amounts in principle to an interference with the free manifestation of religion or belief. The crucial difference with the *forum internum*, however, is that this freedom to manifest one's religion may be limited under strictly prescribed conditions. The *forum externum* 'may be subject only to such limitations as are prescribed by law and are necessary to protect public safety, order, health, or morals or the fundamental rights and freedoms of others'.¹⁰ This test serves to assess the legality, legitimacy, and necessity of the limitations made on the free exercise of religious freedom and, in order to surpass this test, plausible answers must be provided by states: Was the restriction foreseeable, that is, provided by law (legality or rule of law test)? Was the restriction imposed to promote one of the listed goods (legitimacy)? And was the restriction absolutely necessary so as to uphold that good (necessity)?

⁷ International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR); and Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR).

⁸ United Nations Human Rights Committee (UNHRC), 'General Comment No 22: Article 18 (Freedom of Thought, Conscience or Religion)' (1993) UN Doc CCPR/C/21/Rev.1/Add.4, para 2.

⁹ Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR), art 18. See also ICCPR, art 18; ECHR, art 9.

¹⁰ ICCPR, art 18(3). See also ECHR, art 9.

Turning to the AI context, let us consider scenarios when the ‘duty to respect’ religious freedom is breached. The inner freedom, the *forum internum*, would be violated if a state were to ban ‘the belief in AI’. Anthropological fieldwork suggests the existence of religious,¹¹ including theistic, conceptions of AI, ranging from feelings of being ‘blessed by the algorithm’¹² to—more¹³ or less¹⁴—full-fledged AI religious systems revolving around the deification of AI or the belief in the ‘grand narratives’ of AI, such as AGSI (‘Artificial General Super Intelligence’).¹⁵

Accordingly, if a state were to prohibit by law deep AI beliefs such a coercive measure would amount to a straightforward breach of the *forum internum*. As previously explained, as this freedom is absolute there can be no justification for any such interference. This is the case even if the state deems any such beliefs dangerous, a threat to the peace, undermining public safety, public order, or by reference to any other public good.

More academic would be the scenario of a state interfering in the very development of beliefs by AI-operated non-human beings. Since deep AI of the type required for this interference to be even possible does not yet exist, in this instance we enter the realm of science fiction. Still and all, what if the state were to call the robot Cutie from the above-mentioned Asimov story to a halt, hit the kill switch the moment it embarks on its religio-existential quest, and starts developing a religious mindset? Naturally, religious freedom cannot be mobilised against such interference, at least not to the extent it affects the non-human Cutie. As AI has no human conscience and AI is, for the time being, in any event not granted legal personality on the international plane, internationally codified human rights simply do not apply here. For that to change, either the present international legal framework would need to become inclusively interpreted, or AI legal personality and rights-holdership would need to materialise separately.¹⁶ The workings of international policymakers, however, move in precisely the opposite direction; they adamantly insist there can be no such thing as legal personality for AI.¹⁷ There are

¹¹ See Beth Singler, ‘An Introduction to Artificial Intelligence and Religion for the Religious Studies Scholar’ (2017) 20 *Journal of Implicit Religion* 215–32.

¹² Beth Singler, ‘“Blessed by the Algorithm”: Theistic Conceptions of Artificial Intelligence in Online Discourse’ (2020) 3 *AI & Society* 945–55.

¹³ Eg the Turing Church, described as a New Religious Movement (NRM) by Singler (*ibid* section 4 (‘AI new religious movements’)).

¹⁴ One ‘Church of AI’, formally established as the Way of the Future (WOTF) by former Google engineer, Anthony Levandowski, was formally dissolved within a year of its creation.

¹⁵ See Boris Rähme, ‘Artificial Intelligence and Religion: Between Existing AI and Grand Narratives’ (2021) 17(4) *Journal of Objects, Art and Belief* 1.

¹⁶ See the chapter by Klaus Heine in this volume.

¹⁷ Eg European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL)) [2021] OJ C404/107, providing in the Annex at sub-s (6) that: ‘Any required changes in the existing legal framework should start with the clarification that AI-systems have neither legal personality nor human conscience, and that their sole task is to serve humanity’.

dissenters in this debate, but they hint at the benefits of legal personality from an *accountability* perspective,¹⁸ not from a rights-based approach.

But what about the religious freedom of the creators of Cutie, or more generally, of the developers of the (deep) AI techniques—the, say, ‘humans behind the machine’? For the latter’s *forum internum* or *forum externum* to be affected, their religious belief system or practices would need to be tightly connected with the AI techniques in development. The link would need to be inextricable indeed. After all, the state blocking the (further) development of an AI technique would not, as such, stand in the way of the human belief in deep AI. If religious freedom would be claimed with respect to AI techniques and their development, a secular judge would probably be wise to bring virtually any aspect of the *development of AI* under the scope of the *manifestation* of religion, where public good limitations are—contrary to the inner belief freedom—possible in principle. The alternative would render the development of any AI technique immune from scrutiny and restriction the moment the *forum internum* of the creators behind the technique is invoked.

Cases involving the freedom to manifest religious practices that involve (alleged) AI elements, although scarce, do exist. The 1979 European Court of Human Rights (ECtHR) case of *X and the Church of Scientology v Sweden* revolved around a religious artefact called the ‘E-Meter’.¹⁹ Redesigned and patented by the Church of Scientology on numerous occasions, the earliest versions of the E-Meter were, in fact, not invented by Scientology’s founder, L Ron Hubbard (although he did join forces with inventors to bring separate ‘Hubbard’ and ‘Scientology E-Meters’ on the market). Be that as it may, the instrument allegedly serves to measure whether someone’s confession was successful. More specifically, the E-Meter, according to the applicants, the Church of Scientology, and one of its Swedish ministers, is a ‘religious artefact used to measure the state of electrical characteristics of the “static field” surrounding the body and believed to reflect or indicate whether or not the confessing person has been relieved of the spiritual impediment of his sins’.²⁰

The Swedish Ombudsperson took legal action against the manner in which the device was advertised. An injunction to have certain passages redacted was successful before a court and upheld in higher instances. The Church of Scientology, and one of its ministers, complained before the European Commission of Human Rights (ECommHR) that that restriction amounted to a violation of the Church’s freedom of religion and expression. The ECommHR emphasised that the ‘Market Court did not prevent the Church from selling the E-Meter or even advertising it for sale as such. Nor did the Court restrict in any way the acquisition, possession or

¹⁸ Eg Vagelis Papakonstantinou and Paul de Hert, ‘Refusing to Award Legal Personality to AI: Why the European Parliament Got it Wrong’ (*European Law Blog*, 25 November 2020) <<https://europeanlawblog.eu/2020/11/25/refusing-to-award-legal-personality-to-ai-why-the-european-parliament-got-it-wrong/>>.

¹⁹ *X and Church of Scientology v Sweden* App no 7805/77 (ECommHR, 5 May 1979).

²⁰ *ibid* ‘Summary of the Facts’.

use of the E-Meter'.²¹ The ECommHR further reasoned that the right to freedom of religion or belief does not render protection to acts that are of a 'purely commercial nature', like advertising goods for profit, even if this 'may concern religious objects central to a particular need'.²² In sum, the state's actions against the advertisement of the E-Meter—rather than against the production or usage of the E-Meter per se—did not even amount to an 'interference' with the right to freedom of religion or belief.

Now, whether the device central to this case could be deemed an AI instrument *avant-la-lettre* is obviously debatable, however, the relevance of a case like this is that, in practice, the state may decide to intervene with the development or usage of AI(-like) techniques. That is precisely what Sweden did in this case, albeit indirectly, by challenging the advertisement of the technology involved. The harshest interference imaginable would have been the one whereby the state would have altogether banned the production and dissemination of the E-Meter, which was not the case.

The number of examples of AI being deployed by religious groups or churches as part of their religious mission is rapidly increasing. AI can be used for instance:

- in religious studies, including as part of textual analysis of religious sources;²³
- in the process of digitisation and in religious group's social media and wider communication strategies;²⁴

- in the form of robot priests or spiritual machines—in the discharge of core religious rituals, such as religious services, confession, blessings, or prayer.²⁵

Should the state intervene with respect to any of these developments, an interference with the freedom to manifest religion or belief, as practised individually or collectively, is likely to occur. In the future, such cases will likely turn on the necessity test: Was it absolutely necessary for the state to curtail the development, dissemination, and/or usage of the AI-based technique or artefact, in order to protect the opposing value at stake, be it public safety, the rights of other persons, or one of the other recognised grounds for limiting the freedom to manifest one's religion or belief?

Thus far we have focused on the duty to respect religious freedom as conceptualised from the perspective of individual or organised religion. The state may interfere with religious freedom, justified or not, when it imposes restrictions on

²¹ *ibid* 'The Law'.

²² *ibid*.

²³ Eg Mayuri Verma, 'Lexical Analysis of Religious Texts using Text Mining and Machine Learning Tools' (2017) 168 International Journal of Computer Applications 39–45; and Randall Reed, 'AI in Religion, AI for Religion, AI and Religion: Towards a Theory of Religious Studies and Artificial Intelligence' (2021) 12(6) *Religions* 401.

²⁴ Giulia Isetti and others (eds), *Religion in the Age of Digitalization: From New Media to Spiritual Machines* (Routledge 2021).

²⁵ Adrienne Mayor, *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology* (Princeton UP 2018) argues that early attempts at religious automata and artificial life go back to the ancient myths and are less novel or religiously anachronistic than we tend to assume.

AI developments which are incorporated into religious practices. An altogether different angle to the duty to respect religious freedom is the scenario wherein *the state* develops, adopts, or otherwise uses AI technologies with the view towards, or with the result of, limiting the freedom of religion or belief.

For example, AI may be used as part of surveillance techniques to single out persons based on religion, potentially causing discrimination, intimidation, or other forms of ill-treatment of individuals or groups on grounds of religion or belief. Just as public safety measures, in general, may overstep the mark and be overly intrusive, disproportionately affecting privacy or other rights, or be downright discriminatory, the same risk applies in theory to AI techniques incorporated into public safety policies.

Indeed, some such techniques are already in existence and deployed, including in ways that alarmingly illustrate the said risks. Notably, China uses AI techniques to facially profile the Uyghurs, a Turkic ethnic group. The group's religious history is comprehensive and complex, yet present-day Uyghurs compose the second-largest predominantly Muslim group in China. The Uyghurs are under intense surveillance by the Chinese government, facilitated by intrusive surveillance techniques, including AI-based ones, but also in the form of 're-education camps' where an estimated 1 million Uyghurs are being detained. Moreover, the Uyghurs in Xinjiang form somewhat of a Chinese laboratory for testing AI techniques. For instance, AI emotion-detection software has been tested on the Uyghur people, reducing people to statistical pie charts and supposedly unveiling their levels of anxiety. In this way, the very dubious assumption on the part of the Chinese government is that individuals who harbour dissident feelings against the Chinese authorities may be detected.²⁶

As one surveillance watchdog responded to this latest AI addition to China's surveillance apparatus: 'It makes any kind of dissidence potentially impossible and creates true predictability for the government in the behaviour of their citizens. I don't think that Orwell would ever have imagined that a government could be capable of this kind of analysis'.²⁷ In addition to the gross violations of numerous other civil rights, one may argue that these surveillance measures encroach upon the *forum internum*. Even if the results of these surveillance techniques may be questioned—for one thing, who would not display feelings of anxiety in such a state of 1984—China, quite literally, attempts to get into the head and minds of people, seeking to chart their innermost feelings.

²⁶ Jane Wakefield, 'AI Emotion-Detection Software Tested on Uyghurs' (BBC News, 26 May 2021) <www.bbc.com/news/technology-57101248>.

²⁷ ibid quote from BBC interview with the director of IPVM (an internet security and surveillance research group), Conor Healy.

3 The Duty to Protect

Whereas the duty to respect is essentially a negative obligation forcing states to refrain from adverse action, the *duty to protect* requires states to be active and proactive. More specifically, this obligation sees to the relationship between the state and non-state actors as potential offenders, whereas the duty to respect chiefly deals with the relationship between the state and individuals as rights-holders. The Office of the United Nations High Commissioner for Human Rights (OHCHR) defines the obligation to protect as a duty that ‘requires States to protect individuals and groups against human rights abuses’.²⁸ The obligation to protect thus implies that the state under circumstances ought to step in so as to ‘horizontally’ enforce human rights standards. The state, under this duty, must ensure the protection of human rights vis-à-vis other individuals, groups, or miscellaneous non-state entities such as companies or organisations. This duty does not strictly serve to hold those non-state actors accountable for human rights breaches. Whereas the latter may be possible in some jurisdictions under the dictates of domestic (constitutional and other) law, under international human rights law (IHRL) states are the chief duty bearers. States ratify international human rights treaties and are the principal legal persons that may be held accountable. The relevance of the duty to protect, in all this, is hence that the state’s accountability may significantly be broadened up beyond direct state-instigated breaches of IHRL. That is, under circumstances the state may be held accountable not as a result of breaching human rights standards itself, but as a result of not sufficiently discharging the duty to secure protection of human rights vis-à-vis other individuals or groups.

Zooming in on the specific context of AI and religious freedom, examples that come to mind include the state’s duty to protect religious groups and individual religious practitioners from acts of discrimination and violence and from incitement to such adverse actions. It is particularly in the latter context, that of incitement, more specifically, online advocacy of hatred that incites discrimination or violence, that AI plays a role. Like most discussions concerning AI and human rights, we are dealing with a double-edged sword here. For AI may facilitate the state’s investigation into cybercrimes like online hateful incitement based on religion, among other grounds. And at the same time, recent incidents of mob violence, notably the attacks on the US Congress by Trump supporters, show how AI in the form of algorithms help forge information vacuums where clicks on certain (false) information triggers further information along the same lines, seemingly corroborating the original (false) information from further angles, but actually creating an algorithm-steered tunnel vision.

²⁸ OHCHR (n 5).

The duty to respect, in this context, suggests that the state cannot hide behind the fact that these platforms and their advertising models, algorithms, and approaches to suggested further readings are created and maintained by private business actors. At the very least, the state would be obliged to liaise with these non-state actors to discuss ways of avoiding these platforms being used for the dissemination of hateful incitement. The duty to respect may, under certain circumstances, imply the state needing to go as far as actively penalising companies for not sufficiently combatting hateful incitement. Hence, the state may incur accountability for the role of these private platforms if it does not sufficiently discharge these positive obligations of damage prevention and damage control.

The European Union (EU) has also assumed a leading role as far as its territory is concerned. It does so in a twofold fashion. An attempt at discharging the duty to protect in this area in the actual sense can be seen in a provision like article 6 of the Audiovisual Media Services Directive which stipulates that ‘Member States shall ensure by appropriate means that audiovisual media services provided by media service providers under their jurisdiction do not contain any incitement to hatred based on race, sex, religion or nationality’.²⁹ The 2018 amended and updated Audiovisual Media Services Directive moves beyond pure state accountability towards shared responsibility with the platform providers themselves and proposes self-regulatory duties. While the latter duties would strictly not qualify under the (state-centred notion of the) duty to protect, the fact that such duties are imposed by public authority gives the regulatory endeavour a proactive and protective signature. Moreover, these responsibilities are distilled and underscored with a view towards protecting vulnerable groups like children and the general public against harmful content. The relevant 2018 amendment reads in full:

A significant share of the content provided on video-sharing platform services is not under the editorial responsibility of the video-sharing platform provider. However, those providers typically determine the organisation of the content, namely programmes, user-generated videos and audiovisual commercial communications, including by automatic means or algorithms. Therefore, those providers should be required to take appropriate measures to protect minors from content that may impair their physical, mental or moral development. They should also be required to take appropriate measures to protect the general public from content that contains incitement to violence or hatred directed against a group or a member of a group on any of the grounds referred to in Article 21 of

²⁹ Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services [2010] OJ L 95/1 (Audiovisual Media Services Directive), art 6.

the Charter of Fundamental Rights of the European Union (the ‘Charter’), or the dissemination of which constitutes a criminal offence under Union law.³⁰

In addition, outside the realm of audiovisual, media, and other online platforms the duty to respect may force states to act vis-à-vis non-state actors. Within this general area of *business and human rights*, the state’s obligations may be twofold: (i) to liaise with (private) developers of AI techniques to ensure that these techniques are optimally attuned to respect for human rights within the future operational (public service) areas; and (ii) to guard against AI techniques being developed that can be used at the detriment of religious freedom rights.

With respect to the first point, it is not far-fetched to assume that AI will be used within such areas as security, health, and education. Accordingly, AI applications may be used in hospitals, prisons, and schools. In some—if not all—of these areas, the state has express functions, competencies, and human rights obligations. Thus, where health bots may well help discharge the state’s *duty to fulfil* the right to health, at the same time, the state remains—whether under the duty to respect or the duty to protect (in private hospitals)—liable under the duty to respect patients’ freedom of religion or belief.³¹ Similar scenarios may be anticipated in the area of schooling or prison safety and management, among many other areas. Consequently, respect for religious diversity and needs—including dietary ones—is best flagged during the initial design phase of new AI techniques that may become operational within important public and private service areas. The state’s role, there, is a delicate one, hovering between funding and facilitating promising novel techniques that help promote public values such as health, safety, and education whilst liaising with developers to guarantee human rights standards.

More than liaising is clearly needed with respect to the second point: AI techniques may very well be, and indeed already are being, developed and used at the detriment of human rights, including religious freedom. The transnational nature of human rights abuses complicates questions of accountability and enforcement. For instance, if AI surveillance techniques developed by a company adversely impact the privacy and equal rights of a group within the state, the state is to guarantee, under the duty to protect, the rights of this group, failing which it may be responsible for the discriminatory usages of these novel techniques. This is a different scenario from, for instance, Dutch companies that co-develop the AI techniques with which the Chinese government discriminatorily target the Uyghur

³⁰ Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in View of changing market realities [2018] OJ L303/69, recital 47.

³¹ For a more comprehensive analysis, see the chapter by Enrique Santamaría Echeverría in this volume.

community.³² Naturally, China is responsible for the acts of repression, but to what extent are the Netherlands or the EU also responsible?: specifically, responsible for not preventing the exported techniques from being abused to encroach upon human rights?

While Chinese companies may have no scruples about expressly mentioning the profiling potential of novel techniques when filing patents (including express references to Uyghurs),³³ Western companies tend to patent their techniques as abstract technological advancements. These advancements, in any event, are typically sold as but a cog in the wheel of larger surveillance operations. The implicit defence, hence, is that it is not the technique that is at fault, but the user. Under the EU's proposed 'Artificial Intelligence Act' Regulation, any such risks to the enjoyment of fundamental rights are to be assessed and compliance is to be ensured on the basis of a risk classification and human rights compliance scheme.³⁴ However, this pending legislation is largely preoccupied with the EU's internal market.³⁵ Dual-use items—techniques that can be used both for civilian and military purposes—produced within the EU are singled out and subject to export control schemes.³⁶ All in all, this leaves a grey and vulnerable area. Namely, those AI techniques that may be developed and exported and that are on their face but a small element in the larger wheel of abusive measures as used by third countries, thus threatening the fundamental rights beyond the realm of the EU.

Thus far we have focused on AI developed by non-state actors and potentially deployed to the detriment of human rights. A dissimilar area, wherein the duty to protect may be engaged, is the scenario of fake AI techniques. Consumers may be misled by false claims as to the usefulness and benefits of developed AI techniques.³⁷ In the area of religion, we may refer back to the early AI-like case revolving around the E-Meter (see section 2). While consumer protection generally is a

³² Maurits Martijn, 'Berucht Chinees veiligheidsministerie gebruikt Nederlandse software die emoties leest' (*De Correspondent*, 12 July 2019) <<https://decorrespondent.nl/10307/berucht-chinees-veiligheidsministerie-gebruikt-nederlandse-software-die-emoties-leest/317002092-cae75d58>>.

³³ Several 2018 and 2019 patent registrations by Huawei, SenseTime, and Megvii explicitly mention Uyghurs in the context of recognition techniques. See Mary Vlaskamp, 'Onderzoeksureau: Chinese techbedrijven bezig met etnisch profileren, Oeigoeren zijn doelwit' (*De Volkskrant*, 13 January 2021) <www.volkskrant.nl/nieuws-achtergrond/onderzoeksureau-chinese-techbedrijven-bezig-met-etni sch-profileren-oeigoeren-zijn-doelwit~bc8f6317/>.

³⁴ European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

³⁵ Article 2 on scope provides: 'This Regulation applies to: (a) providers placing on the market or putting into service AI systems in the Union, irrespective of whether those providers are established within the Union or in a third country; (b) users of AI systems located within the Union; (c) providers and users of AI systems that are located in a third country, where the output produced by the system is used in the Union.'

³⁶ Regulation (EU) 2021/821 of the European Parliament and of the Council of 20 May 2021 setting up a Union regime for the control of exports, brokering, technical assistance, transit and transfer of dual-use items [2021] OJ L206/1.

³⁷ See the chapter by Shu Li, Béatrice Schütte, and Lotta Majewski in this volume.

pertinent ground for the state to interfere in cases of malfunctioning goods that do not live up to their advertised qualities, the area of religious artefacts is more complex. Here, the state would need to walk a fine line between discharging its duty to protect persons against gross instances of fraud and respect for religious autonomy, that is, the collective realm of religious belief systems wherein a religious community, as guided by its religious leaders, decide over rituals and the meaningful place therein of religious artefacts. One extreme is the scenario wherein the trust of impressionable believers is grossly abused by unscrupulous business churches with a view towards cajoling them out of their hard-earned savings. Under those circumstances, the state's duty to protect would be engaged. Yet, the secular state is not to second-guess every claim of organised religion concerning the alleged powers and benefits of religious artefacts and there is no reason why that consideration would not also extend to the area of (future) AI-based artefacts or AI-driven rituals.

In *X and Church of Scientology v Sweden*, the state had not forced the removal of the E-Meter from the market, but in the interest of consumer protection ruled that some of the terms of the E-Meter's advertisement needed to be redacted. In its periodical, the Church had placed this ad:

Scientology technology of today *demands* that you have your own E-Meter. The E-Meter (Hebbard [*sic*] Electrometer) is an electronic instrument for measuring the mental state of an individual and changes of the state. *There exists no way to clear without an E-Meter.* Price: 850 CR. For international members 20% discount: 780 CR.³⁸

The emphasised parts combine a strong appeal to an alleged religious imperative with a categorical necessity of the instrument. The advertisement's claim, hence, is that 'good believers' must use this device; those refusing to incur the expense may well be doomed. The Swedish Market Court had deemed the formulation misleading and that the consumer—primarily consisting of the readers of the Church's periodical—would be 'particularly susceptible to selling arguments'.³⁹ The ECommHR concluded that:

[T]he Market Court did not prohibit the applicants from advertising the E-Meter and did not issue the injunction under penalty of a fine. The Court chose what would appear to be the least restrictive measure open to it, namely the prohibition of a certain wording in the advertisements. Consequently, the Commission cannot find that the injunction against the applicants was disproportionate to the aim of consumer protection pursued.⁴⁰

³⁸ *X and Church of Scientology v Sweden* App no 7805/77 (ECommHR, 5 May 1979), Summary of the Facts (emphasis added).

³⁹ *ibid* para 5.

⁴⁰ *ibid*.

It may be argued that Sweden's interference with Scientology's E-Meter navigated the tightrope between consumer protection and church autonomy rather well.

4 Duty to Fulfil

The duty to fulfil is arguably the hardest to conceptualise in the present context. Generally, this obligation means 'that States must take positive action to facilitate the enjoyment of basic human rights'.⁴¹ In General Comments on economic, social, and cultural rights, this duty tends to be broken down into three complementary obligations: the obligations to facilitate, promote, and provide.⁴²

The facilitation duty is problematic in the present context, since state involvement with religious affairs may all too easily encroach upon religious autonomy. For instance, AI could—in theory—play a fulfilling role in the seating and security arrangements around major religious events. However, if this duty is allocated to the state as a matter of principle, immediate questions of privacy and general matters of state interference with religious autonomy may arise.

The promotion of religious freedom revolves around important informational and transparency duties. Within the classical religious freedom area, one could think of the duty to inform parents as to the existence of certain opt-outs from religious instruction within public schools. Translated to the AI context, one may recall the health bot from previous discussions and distil a duty to inform patients about certain medical procedures that may encroach on religious convictions so that the patient is in a position to withdraw or seek alternative treatment. In this instance, one sees how the duty to fulfil directly impacts the duty to respect.

Taken at face value, the 'obligation to provide' sounds like an impossibility in the present context—how can the state directly 'provide' religious freedom? However, in analytical terms, the duty signifies for instance in the area of asylum. In that context, states in extreme cases guarantee the right to life of the religiously persecuted person. In a completely different area, the duty has also been mobilised with respect to the repatriation of religious artefacts.⁴³

In the future, AI may—in theory—play a role in those existing, highly specialised, obligation to fulfil areas, or—more likely—broach new ones.

⁴¹ OHCHR (n 5).

⁴² Eg Committee on Economic, Social and Cultural Rights, 'General Comment 15: The Right to Water (UN Doc E/C.12/2002/11, 2002), paras 25–29.

⁴³ For a comprehensive account, see Vanessa Tünsmeyer, *Repatriation of Sacred Indigenous Cultural Heritage and the Law: Lessons from the United States and Canada* (Springer 2021).

5 Conclusion

The triple human rights obligations terminology meaningfully unveils the dynamics between AI and the right to freedom of religion or belief. Especially under the duty to respect and the duty to protect, the widely reported double-edged sword function of AI is visible. AI techniques, also developed by non-state actors, may be used by state and private actors at the detriment of religious freedom, thus engaging crucial human rights obligations in this area. What makes the thematic focus of religious freedom particularly fascinating and complex is that freedom claims may affect the very phenomenon of AI itself (a profound metaphysical belief in AI) as well as the development of AI techniques (AI techniques as part of a religious manifestation). While there is little risk in respecting *forum internum* claims in relation to AI, secular judges will likely push back vis-à-vis any religious freedom claims that relate to the development of AI techniques, lest such development becomes immune from scrutiny.

5

Artificial Intelligence and Freedom of Expression

Giovanni De Gregorio and Pietro Dunn

1 Introduction

Freedom of expression represents a key value to promote a culture of democracy and human rights. Being able to express one's personal ideas and personal opinions, as well as to decide how to convey those ideas and thoughts, allows individuals to fully participate in social life and to project outside values and personality. The Universal Declaration of Human Rights (UDHR),¹ the International Covenant on Civil and Political Rights (ICCPR),² and regional systems—such as the European Convention on Human Rights (ECHR)³—are only some of the instruments that protect freedom of expression as a human right, thus recognising its universal dimension.

The protection of freedom of expression has been deeply impacted by the development of digital technologies and, in particular, by the spread of artificial intelligence (AI) systems. On the one hand, these systems play a critical role in organising large amounts of content or providing opportunities for creativity. On the other hand, this flourishing democratic framework driven by digital technologies does not fully compensate the troubling evolution of the algorithmic society where algorithmic technologies contribute to shaping the definition of online speech and the governance of these expressions is increasingly mediated by AI systems implemented by states and business actors such as in the case of content moderation.

This chapter addresses how freedom of expression and information has been affected by AI technologies in the digital environment. Section 2 explores the main characters and specificities of freedom of expression in the age of AI and focuses on the relationship between AI and the media. Section 3 addresses the private governance of freedom of expression, particularly the use of automated technologies to

¹ Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR), art 19.

² International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR), art 19.

³ Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR), art 10.

moderate online content. Section 4 underscores how the new digital context may require state obligations to ensure the protection of freedom of expression, particularly considering the power exercised by private actors through the implementation of AI systems.

2 Freedom of Expression in the Age of Artificial Intelligence

The rise and consolidation of digital technologies have deeply transformed the way individuals experience and enjoy freedom of expression as a human right.⁴ The spread of AI technologies influences not only the sharing of ideas ('active dimension'), but also the individual freedom to form an opinion based on accessing diverse information ('passive dimension'), as underlined, for instance, by article 10 of the ECHR and confirmed by article 11 of the Charter of Fundamental Rights of the European Union. Indeed, the ECHR, as opposed for instance to the United States' Constitution's First Amendment, provides that freedom of expression encompasses not only the 'freedom to hold opinions' but also that 'to receive and impart information and ideas'. The express reference to both verbs 'receive' and 'impart' reveals how freedom of expression within the traditional constitutional framework of Europe consists of, at least, two equivalent rights: (i) that of the speaker to be allowed to speak and (ii) that of the audience not to be deprived of useful and important information. The latter perspective, which is quite different from that of the United States (US) where, based on the US Constitution's First Amendment, the focus is on the active right of everyone to speak freely, may be translated into a 'right to correct information'.⁵

Artificial intelligence technologies represent an asset for freedom of expression and information. These systems can help increase the possibilities to detect online harmful content and the discoverability of content as well as create further opportunities for media pluralism, thus enriching the informational environment itself. The latest technologies in the field of AI also offer unprecedented tools for artistic, literary, and satirical works—think, for instance, of deepfake technologies which,

⁴ Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 UC Davis Law Review 1149.

⁵ Giovanni Pitruzzella and Oreste Pollicino, *Disinformation and Hate Speech* (Bocconi UP 2020) 91. The authors argue: 'It is entirely evident that the existence of explicit or in any event consolidated constitutional coverage for freedom to be informed argues in favor of mechanisms that allow for "filtering" the mare magnum of the web from content incapable of providing a contribution to information because it is entirely false or groundless or because it is not qualified or verified. If European constitutionalism considers freedom of expression as functional to the informational needs of the public, and thus indirectly to the formation of public opinion and the functioning of democracy, the recognition of the right to be informed translates into providing for the right to correct information'. However, as underscored by the same authors, this approach is typically European, indeed, 'none of this is found in the United States system.'

although being potentially highly risky,⁶ may nonetheless be used to produce unprecedented works of art. The media can also avail itself of unprecedented tools for the production of information.⁷ In fact, the creative potential of AI and algorithms is such that scholars have discussed whether automatically generated content should be considered as a form of speech worthy of constitutional protection.⁸

Nonetheless, the implementation of AI technologies raises questions about the protection of the right to freedom of expression. Algorithmic tools are used more and more frequently for the purposes of governing speech in the digital age, with some direct effects on the way individuals may enjoy freedom of expression. These effects include, most notably, the use of AI for filtering of unwarranted or illegal content, as well as for organising the way content is displayed and presented to the public. Such systems can, on the one hand, make significant mistakes when moderating content, as well as replicate discriminatory approaches and human bias. The civic space where freedom of expression may be experienced and realised is increasingly subject to the mediation and influence of a multitude of AI applications, ‘from newsfeed algorithms to connected devices in smart cities’,⁹ a mediation which, additionally, is often controversial and lacks transparency.

Freedom of expression can also be impaired as a secondary effect of the infringement of other fundamental rights. This happens primarily when freedom of expression intersects with values such as the protection of privacy and equality. Indeed, according to David Kaye, as Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, the use of automated tools for these purposes can deeply affect a range of fundamental rights of individuals around the globe, as well as the rights to privacy and non-discrimination.¹⁰ In fact, in the context of the digital age, the field of freedom of expression and that of privacy and data protection, which initially moved on parallel (but separate) tracks, have been subjected to a progressive process of convergence, especially because of the increased use of AI in the governance of the digital space.¹¹ In order to

⁶ Danielle K Citron and Robert Chesney, ‘Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security’ (2019) 107 California Law Review 1753. For the negative impact on women’s rights, see also the chapter by Marília Papaléo Gagliardi in this volume.

⁷ Natali Helberger and others, ‘Implications of AI-Driven Tools in the Media for Freedom of Expression’ (Council of Europe, 2020) <<https://rm.coe.int/cyprus-2020-ai-and-freedom-of-expression/168097fa82>>.

⁸ Stuart Minor Benjamin, ‘Algorithms and Speech’ (2013) 161 University of Pennsylvania Law Review 1445; Alan M Sears, ‘Algorithmic Speech and Freedom of Expression’ (2020) 53 Vanderbilt Journal of Transnational Law 1327; Manasvin Goswami, ‘Algorithms and Freedom of Expression’ in Woodrow Barfield (ed), *The Cambridge Handbook of the Law of Algorithms* (CUP 2020) 558.

⁹ Privacy International and Article 19, ‘Privacy and Freedom of Expression In the Age of Artificial Intelligence’ (Article 19, April 2022) <www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf>.

¹⁰ David Kaye, ‘Promotion and Protection of the Right to Freedom of Opinion and Expression’ (United Nations 2018) A/73/348 <<https://www.ohchr.org/en/calls-for-input/report-artificial-intelligence-technologies-and-implications-freedom-expression-and>>.

¹¹ Giovanni De Gregorio, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (CUP 2022) 168. As highlighted by the author, cases such as Case C-101/01 Lindqvist, E.C.R. 2003 I-12971 (6 November 2003) and Case C-131/12 Google Spain SL and Google Inc

function properly, those algorithmic tools rely on the use of an extensive quantity of data concerning the identities, interests, and tastes of individual users. However, the collection of individual personal data in the context of the developing surveillance capitalism can represent a strong deterrent for the full enjoyment of freedom of expression.¹²

The collection and storage of information about the identity of individuals, as well as of their thoughts, opinions, ideas, and interests (including sensitive data concerning, for example, the ethnicity or sexual orientation of a person) may, in fact, lead them to hold back from fully enjoying their freedom of expression and information. This is also why anonymity and encryption play a critical role for the promotion of free speech,¹³ by limiting behaviours that would result from the awareness of a surveillance regime.¹⁴

The use of AI has not only transformed the way single individual users experience freedom of expression in the digital age but has also consequences with respect to the specific facet of free speech represented by freedom of the press, with an evident impact both on the media's freedom to impart information and on users' freedom to seek and receive high quality and pluralistic information. Media and journalistic sources today face the specific characteristics of the digital space. A 2020 report by the UNESCO Institute for Information Technologies in Education highlighted some of the main changes that have affected the media environment: (i) the data explosion, that is, the ever-increasing amount of information growing exponentially; (ii) the source and verification ambiguity caused by the reduced role of traditional professional media as gatekeepers of information; (iii) the diminishing power of regulatory authorities vis-à-vis the self-imposition of private digital platforms; (iv) the wide geographical spread of content; (v) the ambiguity in tracing a certain content to its creator or distributor; (vi) the creation of distributed collections of data versus the traditional collection of information being reserved to specific physical places such as libraries, museums, and archives; (vii) media convergence, that is, the erosion of the traditional barriers between the various forms of information production; (viii) the proliferation of devices and

¹² Agencia Española de Protección de Datos (AEPD) and Mario Costeja González, ECLI:EU:C:2014:317 (13 May 2014) have shown that such a convergence can lead to tensions and to an adversarial dialectic between the rights to privacy and data protection and that to freedom of expression. However, these two constitutional values are more and more interrelated through a cooperative relationship in the context of the promotion of a framework of digital constitutionalism in Europe.

¹³ Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power* (Profile Books 2019).

¹⁴ Dirk Voorhoof, 'Internet and the Right of Anonymity' (Proceedings of the Conference Regulating the Internet, Center for Internet Development, 2011) <<http://hdl.handle.net/1854/LU-2045023>>; Jason A Martin and Anthony L Fargo, 'Anonymity as a Legal Right: Where and Why It Matters' (2015) 16 North Carolina Journal of Law and Technology 311.

¹⁴ Privacy International and Article 19 (n 9) 8.

platforms where users may publish or look for content; and (ix) the personalisation and customisation of content.¹⁵

The media and the press hold a special place in the context of freedom of expression since they play an important role as ‘public watchdogs’ by protecting and guaranteeing the correct functioning of democracy.¹⁶ In exercising such a duty, they serve mainly three kinds of tasks: (i) they observe and inform the public; (ii) they participate in public life as independent actors through critical comments, advice, advocacy, and the expression of opinions; and (iii) they provide a channel, forum, or platform for extra-media voices or sources.¹⁷ In this sense, by providing the public with diverse, pluralistic, and critical information and points of view, the media and the press are fundamental actors in the enforcement of the passive dimension of freedom of expression, to be intended as the right to be informed.

Whereas, in the past, the opportunities for individuals to publicly express their thoughts and opinions were restricted due to the practical scarcity of means and resources to disseminate them, meaning that professional actors (including newspapers and broadcasters) previously represented the material gatekeepers of information, whereas today’s online platforms and social media give a chance to everyone to speak their mind on the Internet. However, scarcity of means has been substituted by another form of scarcity, that is, that of attention: because of the ever-increasing and ever-expanding flow of information across the Internet, it has become increasingly difficult for users to understand which sources are worthy of their attention and therefore, to orient themselves within the digital environment and to fully enjoy the advantages of media pluralism.¹⁸

As a matter of fact, the (algorithmic) organisation of the digital space and the increased role of the Internet—particularly social media and social network services—for the dissemination and reception of news have compelled the media to change the way they produce and distribute information. For instance, current journalistic sources deal with an informational environment which is not limited to specific time spans but is always active. Additionally, media outlets take into account the functioning of recommender systems and algorithmic content curation (as will be dealt with in section 3). In this sense, concerns have been raised as to the risk of an impact on the quality of information itself, which would be questionable

¹⁵ Tatiana Murovana, ‘Media and Information Literacy and Artificial Intelligence’ in Ibrahim Kushchu and Tuba Demirel (eds), *Artificial Intelligence: Media and Information Literacy, Human Rights and Freedom of Expression* (UNESCO IIITE, TheNextMinds 2020) 36–37 <https://iite.unesco.org/wp-content/uploads/2021/03/AI_MIL_HRs_FoE_2020.pdf>.

¹⁶ See *The Observer and The Guardian v United Kingdom* App No 13585/88 (ECtHR, 26 November 1991), para 59; *Jersild v Denmark* App No 15890/89 (ECtHR, 23 September 1994), para 31.

¹⁷ Helberger and others (n 7) 7.

¹⁸ De Gregorio (n 11).

not only from a media freedom perspective, but also from the point of view of users' right to pluralistic and quality information.¹⁹

Besides, media convergence and the role of personalisation and customisation of content have seemingly impacted on the way the media and the press produce and disseminate information. On the one hand, the advent of social media and social networks has given the media a new essential platform to distribute multi-media content capable of reaching a large audience. On the other hand, this migration of the information market to the Internet has had to face the typical goal of algorithmic systems for content curation which is that of maximising users' engagement and retention through the creation of a digital space constructed especially for them, that is what Sunstein effectively described as the 'Daily Me'.²⁰ As has been pointed out, this could, nonetheless, have a detrimental effect on the quality of journalistic sources—and of the media and the press in general—due to the need to adjust to the algorithms created by private oligopolists governing the Internet. Indeed, 'the advertising-driven business models at the core of today's internet structure have profoundly affected the sustainability of legacy media by structurally shifting power, to the detriment of quality journalism', and 'the use of AI technologies further shifts this imbalance—with a particular impact in countries with low internet penetration or no strong public service media'.²¹

In addition, the creation of a 'Daily Me' also impacts the other side of the coin, that is, people's right to receive pluralistic information. In fact, users of the Internet generally end up being locked within echo chambers and filter bubbles²² where they have limited opportunity of being exposed to diverse content and diverse points of view. In turn, this can also lead to serious effects and consequences, such as the polarisation of the political debate, thus contributing to the rise of digital populist narratives.²³

These challenges are primarily connected to the private governance of freedom of expression in the digital age. Contemporary speech gatekeepers, notably social media as well as other Internet intermediaries such as search engines, have come to rely more and more on the use of AI for the purposes of governing the services and digital spaces they offer. Since digital infrastructures have become the new avenues for the exercise of freedom of expression and for the dissemination of individual

¹⁹ Centre for Media Transition, 'The Impact of Digital Platforms on News and Journalistic Content' (*Australian Competition & Consumer Commission*, 2018) <<https://www.uts.edu.au/node/247996/projcts-and-research/impact-digital-platforms-news-and-journalistic-content>>.

²⁰ Cass R Sunstein, *#Republic: Divided Democracy in the Age of Social Media* (Princeton UP 2017).

²¹ Julia Haas, 'Freedom of the Media and Artificial Intelligence' (*Global Conference for Media Freedom*, 16 November 2020) 2–3 <www.international.gc.ca/world-monde/assets/pdfs/issues_development-enjeux_developpement/human_rights-droits_homme/policy-orientation-ai-ia-en.pdf>.

²² Eli Pariser, *The Filter Bubble: What the Internet is Hiding From You* (Penguin 2011); Sunstein (n 20); Frank Pasquale, *New Laws of Robotics: Defending Human Expertise in the Age of AI* (Belknap Press of Harvard UP 2020).

²³ Oreste Pollicino and Giovanni De Gregorio, 'Constitutional Law in the Algorithmic Society' in Amnon Reichman and others (eds), *Constitutional Challenges in the Algorithmic Society* (CUP 2021) 3.

thoughts and opinions, the private owners of those digital infrastructures are, nowadays, key components within the governance of contemporary speech. The role of online platforms, particularly social media, in defining the standards of free speech online constitutes one of the primary questions for the protection of the right to freedom of expression.

3 The Private Governance of Freedom of Expression

The private governance of online speech is another critical area for freedom of expression in the age of AI. In particular, the discretion of online platforms as private actors in defining the rules for the moderation of content, or of online speech, raises questions about the protection of this human right. New private actors have become central within the power dynamics of speech governance²⁴ and have come to play the role of the new gatekeepers of information.²⁵

Because of the extraordinary amount of content produced and posted online each day, providers of intermediary services, and online platforms in particular, currently make ample resort to AI for the purposes of content moderation and curation.

Grimmelmann defines content moderation, *lato sensu*, as ‘the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse’,²⁶ and thus includes within it a set of different practices and activities ranging from the exclusion of unwanted members to pricing to the organisation of content to setting of norms (that is, terms and conditions). Such a definition, in fact, includes two different facets of content moderation. On the one hand, ‘hard moderation’²⁷ (or content moderation *stricto sensu*) focuses on the removal of user-generated content that is either illegal or non-compliant with the platform’s terms and conditions. Possible reactions may include, notably, the removal of unwarranted content, or even the suspension or deletion of the user’s account.²⁸ On the other hand, ‘soft moderation’,²⁹ or ‘content

²⁴ Jack M Balkin, ‘Free Speech Is a Triangle’ (2018) 118 Columbia Law Review 2011.

²⁵ Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale UP 2018).

²⁶ James Grimmelmann, ‘The Virtues of Moderation’ (2015) 17 Yale Journal of Law and Technology 42, 47.

²⁷ Robert Gorwa, Reuben Binns, and Christian Katzenbach, ‘Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance’ (2020) 7 Big Data and Society 1, 3.

²⁸ Kate Klonick has classified the various forms of content moderation, distinguishing in particular between ex ante and ex post moderation, depending on whether control is operated before or after the uploading of the content to the platform. Ex post moderation, moreover, can be either proactive, meaning that the relevant intermediary actively looks for items which must be taken down, or reactive, meaning that it responds to notices and flags by other users of the service. See Kate Klonick, ‘The New Governors: The People, Rules, and Processes Governing Online Speech’ (2017) 131 Harvard Law Review 1598.

²⁹ Gorwa, Binns, and Katzenbach (n 27) 3.

curation,³⁰ consists of the operational decisions concerning the organisation, display, presentation, and suggestion of content across the Internet.³¹ As underscored by Gillespie, content moderation is, in practice, ‘the commodity that platforms offer’,³² since it serves, first and foremost, the purpose of ensuring that users can enjoy a positive experience of their services.³³

Hybrid moderation systems, combining the action of humans and automated decision-making (ADM) systems, are especially popular, with algorithms operating a pre-emptive classification of user-generated content.³⁴ Apart from guaranteeing a more efficient work of content moderation, this is essential to improve the working conditions and quality of life of human reviewers, who are in most cases at risk of burnout, desensitisation, and even PTSD (‘Post-Traumatic Stress Disorder’) because of the repeated and continuous exposition to harmful material.³⁵

As regards curation, recommender systems are the main tools used to suggest relevant items to users and to organise the presentation and dissemination of content. They are, in practice, ‘functions that take information about a user’s preferences (eg about movies) as an input, and output a prediction about the rating that a user would give of the items under evaluation (eg new movies available), and predict how they would rank a set of items individually or as a bundle’.³⁶

The use of AI systems for such tasks raises nonetheless a range of issues and challenges concerning the respect of the individual fundamental right to freedom of expression. Most notably, when it comes to hard moderation, the main issue of ADM systems is represented by the inevitability of mistakes. Artificial intelligence technologies classifying systems fuelled by the analysis of data are based on statistic and probabilistic bases that lead to a certain margin of error. This can either translate into false positives, meaning that a legitimate content is classified as being in violation of the law or of the platform’s terms and conditions, or, vice versa, into false negatives. Moreover, the relationship between the two types of error is

³⁰ Emma Llansó and others, ‘Artificial Intelligence, Content Moderation, and Freedom of Expression’ (TWG 2020) <www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf>.

³¹ In this sense, Wu speaks of ‘positive’, or ‘affirmative’, speech control, highlighting how the practices of content curation can indeed affect the dissemination of posts and other items across the Internet. See Tim Wu, ‘Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems’ (2019) 119 Columbia Law Review 2001, 2014.

³² Gillespie (n 25) 13.

³³ Richard Wilson and Molly Land, ‘Hate Speech on Social Media: Content Moderation in Context’ (2021) 52 Connecticut Law Review 1029, 1054.

³⁴ Klonick (n 28) 1635 ff; Cambridge Consultants, ‘Use of AI in Online Content Moderation’ (Ofcom 2019) <https://www.ofcom.org.uk/_data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf>; Giovanni Sartor and Andrea Loreggia, ‘The Impact of Algorithms for Online Content Filtering or Moderation. “Upload Filters”’ (European Parliament 2020) JURI Committee PE 657.101 <[www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU\(2020\)657101_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf)>, 22.

³⁵ Sarah T Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale UP 2019).

³⁶ Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi, ‘Recommender Systems and Their Ethical Challenges’ (2020) 35 *AI & Society* 3.

inversely proportional—therefore, for example, a system calibrated to avoid any false negatives will lead to a rise in the rate of false positives.

The challenge is thus that of correctly balancing two often conflicting goals, that is, on the one hand, the reduction of dangerous, harmful, and/or illegal content spread across the Internet and, on the other hand, the protection of the individual right to freedom of expression against the over-removal of online content. The threshold can vary sensitively depending on the type of content which the system seeks to moderate: filtering tools for copyright enforcement may have to face challenges that are quite different from those posed by hate speech or disinformation detectors. In the case of hate speech, for instance, a serious issue is represented by the risk of discriminatory outputs silencing discriminated or marginalised communities and groups: indeed, in many cases, classification systems are not capable of recognising the language patterns that are typical of certain categories of people (eg African American English, as well as the jargon of LGBTQI+ people) and thus end up misinterpreting the meaning of certain words, terms, and expressions.³⁷

Artificial intelligence systems for content moderation are, indeed, generally built on databases and corpora that mainly (if not exclusively) focus on mainstream language. This is true not only from a sociological point of view, meaning that minority languages are not considered sufficiently when moderating content in Western languages (eg English), but also from a geographical point of view. Platforms and social media often lack the economic interest in managing and translating their terms and conditions to adapt them to peripheric areas of the world such as Africa or Asia,³⁸ as well as in building efficient technologies capable of applying those terms and conditions or the law. Algorithms for detecting hate speech, for instance, are not at all efficient in understanding when hate speech is uttered in Swahili or Burmese. This has led to some tragic consequences, as happened in 2017 when Facebook's algorithms proved to be unable of recognising and removing incendiary anti-Rohingya hate speech in Burmese and thus contributed to the dramatic Rohingya genocide.³⁹ Large social media companies are thus not capable of ensuring high-quality moderation performances at the local level, with

³⁷ See eg Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber, 'Racial Bias in Hate Speech and Abusive Language Detection Datasets' in Sarah T Roberts and others (eds), *Proceedings of the Third Workshop on Abusive Language Online* (Association for Computational Linguistics 2019) <www.aclweb.org/anthology/W19-3504>; Maarten Sap and others, 'The Risk of Racial Bias in Hate Speech Detection' in Anna Korhonen, David Traum, and Márquez Lluís (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics 2019) <www.aclweb.org/anthology/P19-1163>; Thiago Dias Oliva and others, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2021) 25 Sexuality & Culture 700.

³⁸ Giovanni De Gregorio and Nicole Stremlau, 'Platform Governance at the Periphery: Moderation, Shutdowns and Intervention' in Judit Bayer and others (eds), *Perspectives on Platform Regulation. Concepts and Models of Social Media Governance Across the Globe* (Nomos 2021) 433.

³⁹ Nicolas P Suzor, *Lawless: The Secret Rules That Govern Our Digital Lives* (CUP 2019); Steve Stecklow, 'Why Facebook Is Losing the War on Hate Speech in Myanmar' (Reuters, 15 August 2018) <www.reuters.com/investigates/special-report/myanmar-facebook-hate/>.

consequences that can be problematic for the protection of freedom of expression and information, as well as of public order and diversity.⁴⁰ The silencing of minority voices, which may take place for the limited will of platforms to build algorithms capable of dealing with non-mainstream languages and jargons, represents not only a restriction of speakers' freedom of expression but also a limitation for the wider public's right to seek and receive information from those sources.

Moreover, the inherently business-driven nature of recommender systems has also proven in many cases to actually enhance the risks connected to harmful content, including most notably hate speech and disinformation, that are forms of speech that tend to attract users' engagement. Recommender systems have been found to promote precisely that kind of content,⁴¹ while affecting, as noted above, the discoverability of items published by (and for) minorities and marginalised or discriminated categories of people,⁴² sometimes even leading to forms of 'shadow-banning'.⁴³

These challenges raise questions on the role and responsibilities of online platforms and states to protect human rights in the digital age. Particularly, the private governance of online speech leads to look at business and human rights as well as at state positive obligations to protect human rights.

4 State and Business Obligations to Protect Freedom of Expression

The purpose of international human rights law (IHRL), both at a global and at a regional level, is not only that of identifying and defining the scope of human rights but also that of setting basic rules for states to respect, protect, and fulfil those rights.⁴⁴ In this respect, states deal with negative and positive obligations: whereas

⁴⁰ Cf Yohannes Eneyew Ayalew, 'Uprooting Hate Speech: The Challenging Task of Content Moderation in Ethiopia' (*Open Internet for Democracy*, 27 April 2021) <<https://openinternet.global/news/uprooting-hate-speech-challenging-task-content-moderation-ethiopia>>.

⁴¹ Llansó and others (n 30).

⁴² On the notions of discoverability, see Eleonora Maria Mazzoli and Damian Tambini, 'Prioritisation Uncovered: The Discoverability of Public Interest Content Online' (*Council of Europe*, DGI(2020)19 2020) <<https://rm.coe.int/publication-content-prioritisation-report/1680a07a57>>; Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York UP 2018).

⁴³ Carolina Are, 'The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram' (2022) 22(8) Feminist Media Studies 2002; Gabriel Nicholas, 'Shedding Light on Shadowbanning' (*Center for Democracy and Technology*, April 2022) <<https://cdt.org/insights/shedding-light-on-shadowbanning/>>.

⁴⁴ See Office of the United Nations High Commissioner for Human Rights (OHCHR), 'International Human Rights Law' (*OHCHR*) <www.ohchr.org/en/professionalinterest/pages/internationallaw.aspx>, positing: 'International human rights law lays down obligations which States are bound to respect. By becoming parties to international treaties, States assume obligations and duties under international law to respect, to protect and to fulfil human rights. The obligation to respect means that States must refrain from interfering with or curtailing the enjoyment of human rights. The obligation to protect requires States to protect individuals and groups against human rights abuses. The obligation to fulfil means that States must take positive action to facilitate the enjoyment of basic human rights'.

the former consists of a requirement for states not to interfere in the exercise of rights, the latter consists of the actual obligation to do something, that is, ‘to take the necessary measures to safeguard a right or, more precisely, to take the necessary measures to safeguard a right or, more specifically, to adopt reasonable and suitable measures to protect the rights of the individual’.⁴⁵

In the context of the digital age, the question arises as to the existence of positive obligations of states to actively protect freedom of expression and information vis-à-vis the new challenges raised in the aftermath of the spread of the Internet and the private governance of freedom of expression. As mentioned above, today’s free speech is largely and widely in the hands of private platforms governing the publication and dissemination of content. While social media offer spaces that empower users to access online content, they are also powerful actors that organise that content by implementing algorithmic technologies.

Strategies may include the resort to self- and co-regulatory schemes and/or the sensitisation of private companies to human rights.⁴⁶ States can require the private sector to respect human rights, for instance, by introducing legal measures to restrict or influence the development and implementation of AI applications as well as through the introduction of human rights risk assessments. Within the framework of the Council of Europe (CoE), it has been argued that, consistently with case law from the European Court of Human Rights (ECtHR), contracting states have a duty to ensure that private governance of the digital space, especially AI-driven governance, is not detrimental to the passive and active dimensions of the individual right to freedom of expression and information. Thus, for instance, the CoE has recommended that individuals should be informed about algorithmic decision-making affecting them and have meaningful control over the processes leading to such decisions, including via the possibility of gaining access to effective remedies.⁴⁷ The 2018 Recommendation of the Committee of Ministers of the Council of Europe on the Roles and Responsibilities of Internet Intermediaries clearly moves in such a direction.⁴⁸

When discussing the positive obligations of states with respect to the protection of human rights vis-à-vis private actors, another aspect worthy of attention is represented by the expansion of the scope of action of human rights law (HRL) itself entailed by those obligations.⁴⁹ Indeed, whereas in the past HRL was mainly

⁴⁵ Jean-François Akandji-Kombe, ‘Positive Obligations under the European Convention on Human Rights: A Guide to the Implementation of the European Convention on Human Rights’ (*Council of Europe* 2007) 7 <[www.echr.coe.int/LibraryDocs/DG2/HRHAND/DG2-EN-HRHAND-07\(2007\).pdf](http://www.echr.coe.int/LibraryDocs/DG2/HRHAND/DG2-EN-HRHAND-07(2007).pdf)>.

⁴⁶ Kaye (n 10) 9–10.

⁴⁷ Helberger and others (n 7) 17.

⁴⁸ Council of Europe (CoE), Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries (7 March 2018).

⁴⁹ John H Knox, ‘Horizontal Human Rights Law’ (2008) 102 American Journal of International Law 1. For a comprehensive discussion of business and human rights in the age of AI, see the chapter by Isabel Ebert and Lisa Hsin in this volume.

focused on state activity and state duties, the contemporary framework is more and more aware of the role and responsibilities of private corporations, as well exemplified, for example, by the Guiding Principles on Business and Human Rights. Indeed, states must ‘protect against human rights abuse within their territory and/or jurisdiction by third parties, including business enterprises’ and ‘set out clearly the expectation that all business enterprises domiciled in their territory and/or jurisdiction respect human rights throughout their operations’.⁵⁰

In this sense, states’ positive obligations to protect free speech and related rights translate into the development of a ‘horizontal effect’ of such rights: single individuals should, pursuant to such an effect, be able to enforce their fundamental right to freedom of expression also with respect to private actors. However, the true existence of such a horizontal effect is not at all an established principle at a worldwide level. Most notably, the ‘state action doctrine’ of US constitutional law, that is, the principle pursuant to which the provisions of the US Constitution are only applicable to state actors,⁵¹ represents a hurdle to the enforceability of the First Amendment against private actors, including online platforms, social media, and social networks.⁵²

The matter of the horizontal effect of fundamental human rights is still a debated topic also in Europe,⁵³ as also underlined by the case law of European courts:⁵⁴ among others, the applicability of such a principle to the context of freedom of expression and information is not fully clear.⁵⁵ Nonetheless, as mentioned above,

⁵⁰ Office of the United Nations High Commissioner for Human Rights (OHCHR), ‘Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework’ (HR/PUB/11/04) (OHCHR 2011) 3 <www.ohchr.org/sites/default/files/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf>.

⁵¹ Richard S Kay, ‘The State Action Doctrine, the Public-Private Distinction, and the Independence of Constitutional Law Symposium on the State Action Doctrine’ (1993) 10 Constitutional Commentary 329, 330, arguing: ‘The distinction between public and private manifests itself in several difficult and intensely contested questions of constitutional adjudication. Most directly relevant is the reach of the “state action” doctrine in connection with certain provisions of the Constitution, most notably the Equal Protection and Due Process Clauses of the Fourteenth Amendment. These provisions have been held to apply only to the infliction of injuries that can somehow be attributed to a “state.” The infliction of similar injuries by private persons, under this doctrine, are left unregulated by constitutional rule. Besides, *Marsh v Alabama* 326 US 501 (1946) held that private actors may be required to respect individuals’ free speech rights, pursuant to the First Amendment, whenever they act in lieu of the state with respect to public fora. *Marsh v Alabama*, in particular, concerned the prohibition imposed by a private company town to distribute religious literature upon its territory. In that case, the Supreme Court held that the company town had violated the First Amendment rights of the applicant, a Jehovah’s Witness named Grace Marsh. Some scholars, including Gillespie (n 25), have suggested extending the notion of ‘public fora’ to online platforms as well.

⁵² Jonathan Peters, ‘The “Sovereigns of Cyberspace” and State Action: The First Amendment’s Application—Or Lack Thereof—to Third-Party Platforms’ (2017) 32 Berkeley Technology Law Journal 989. See *Prager University v Google LLC* 951 F.3d 9912 (2020); and *NetChoice v Paxton* 596 US ___ (2022).

⁵³ Eleni Frantziou, ‘The Horizontal Effect of the Charter: Towards an Understanding of Horizontality as a Structural Constitutional Principle’ (2020) 22 Cambridge Yearbook of European Legal Studies 208.

⁵⁴ Oreste Pollicino, *Judicial Protection of Fundamental Rights Online. A Road Towards Digital Constitutionalism?* (Hart 2021).

⁵⁵ Naomi Appelman, João Pedro Quintais, and Ronan Fahy, ‘Article 12 DSA: Will Platforms Be Required to Apply EU Fundamental Rights in Content Moderation Decisions?’ (*DSA Observatory*,

the European framework expressly recognises that freedom of expression encompasses an active and passive dimension. In this sense, freedom of expression is not only directly entailed by the concept of human dignity, but it is also valuable from a collective point of view since it ensures the protection and promotion of the democratic process.

It should not come as a surprise, therefore, that the positive dimension of human rights has led the European Union (EU) to react to threats facing freedom of expression and the private governance of online speech by proposing legal instruments, particularly the Digital Services Act (DSA)⁵⁶ which is intended to be a central piece of the developing European digital constitutionalism.⁵⁷ Overall, the DSA includes a range of provisions aimed at setting some limitations on the power of online platforms. For instance, the DSA introduces relevant duties of transparency as regards the use of recommender systems and the micro-targeting of users for advertising purposes,⁵⁸ as well as an obligation for online platforms to inform users about any content moderation or curation action taken against them and about the reasons that led to such action,⁵⁹ and to give them the means to propose a complaint against such decisions.⁶⁰

However, when moving to other areas of the world, the situation concerning the protection of freedom of expression at the intersection of public and private powers appears to be more complex. There is still a significant gap between the way online platforms and social media deploy their content moderation practices in the US and in Europe as opposed to Africa or Asia, most notably as a result of the limited technological resources to develop AI technologies that can deal with non-Western languages. Apart from the above-mentioned issues concerning the enhanced risk for discrimination and errors, the incapability of private actors to recognise and remove illegal and harmful content, including hate speech and disinformation, has led many countries to react and find justifications to adopt highly restrictive measures (eg Singapore, Malaysia, and Nigeria) or opting to impose internet shutdowns. Shutdowns, in particular, can consist of various strategies such as the slowing down of the Internet (with a view to making it practically unusable) or the more drastic choice of switching it off entirely.⁶¹ Evidently, such reactions constitute a serious infringement of human rights connected to freedom

⁵¹ May 2021) <<https://dsa-observatory.eu/2021/05/31/article-12-dsa-will-platforms-be-required-to-apply-eu-fundamental-rights-in-content-moderation-decisions/>>.

⁵⁶ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277/1.

⁵⁷ De Gregorio (n 11).

⁵⁸ DSA, arts 27, 38.

⁵⁹ DSA, art 17.

⁶⁰ DSA, art 20.

⁶¹ De Gregorio and Stremlau (n 38).

of expression, showing how governments may have limited means to address the challenges raised by the spread of harmful content. Additionally, they demonstrate once more that the way platforms and social media govern online speech, also through the use of AI, can have highly relevant impacts on the way free speech is enjoyed by individuals globally, particularly in the case of harmful content, such as hate speech,⁶² that raises questions beyond the local dimension and to the international community.

5 Conclusion

In the digital age, the right to freedom of expression has been highly impacted by the development and deployment of AI systems. Although these technologies represent an extraordinary tool for the contemporary production of and control over online speech, they also raise a range of significant issues and challenges from numerous points of view.

First, the use of AI in the organisation of content in the digital environment has affected the way users access information and the media produce and distribute information, sometimes to the detriment of quality information. A result which could—with the media being the ‘public watchdogs’ of democracies—ultimately be detrimental to society as a whole. The influence that AI exerts on freedom of expression and information is particularly strong as it is grounded on the collection of personal data for the purposes of profiling, micro-targeting, and customisation. Besides, this business model, which is a direct reflection of surveillance capitalism, could also represent a deterrent to the free expression of one’s thoughts and opinions: knowing that our personal information is being collected and that our words and opinions may be easily tracked back to us could lead us to limit ourselves in expressing them. This shows how much, in the context of the algorithmic society, freedom of expression and information is strictly intertwined with other fundamental rights, notably privacy and data protection.

Second, automated means of content moderation and curation can seriously impair the full enjoyment of individual freedom of expression in its active dimension, both because high error rates could lead to the over-removal of content and because recommender systems could excessively reduce the visibility and discoverability of specific items, sometimes leading to proper forms of shadow-banning. Automated content moderation and curation are especially problematic when it comes to dealing with languages and jargon that are not mainstream, both socio-logically and geographically, and that, as such, are not sufficiently represented within databases and corpora used for the training of algorithms: as a result,

⁶² ibid.

moderation of content produced by minority groups, such as ethnic subgroups of the population or members of the LGBTQI+ community,⁶³ or in languages that are less diffused globally, is often affected by much higher error rates (both in terms of false negatives and in terms of false positives).

Third, it must be stressed that the resort to those AI systems is generally operated by private actors who have become the new protagonists of speech governance dynamics. The increased role of corporations in this area calls for increased human rights responsibilities and requires states to actively engage in positive obligations to foster the protection of freedom of expression and information in the context of the digital age. States have only adopted preliminary steps with respect to the need of containing the one-sided power of digital private actors, nor does state intervention seem viable in all jurisdictions despite the universal nature of human rights, due to constitutional divergences across the globe. Moreover, the new dynamics of power in the digital sphere as far as freedom of expression is concerned cause significant issues in a range of African and Asian countries where social media companies and platforms have limited interest in developing AI systems efficiently adapted to the local linguistic, cultural, or political conditions.

The challenges to freedom of expression brought forward by the deployment of AI applications underline how these systems raise questions about unwarranted and excessive impact upon this human right, both vis-à-vis public powers, as well as the new private ones. Although there are some attempts to address this situation, as suggested, for instance, by the proposal and adoption procedure of the DSA in the context of the EU, legal measures to deal with the challenges raised by AI technologies to freedom of expression are still, at a global level, in their infancy.

⁶³ See also the chapter by Masuma Shahid in this volume.

6

Artificial Intelligence and Freedom of Assembly

Margaret Warthon

1 Introduction

As more protests against the Russian regime erupt, Putin's regime is flooding the country with surveillance technology. Moscow alone has some 214,000 CCTV cameras, some with live facial recognition technology (FRT).¹ In addition, there are plans underway to develop 'aggression detection'² and 'smart' riot surveillance systems based on FRT.³ The developers of these technologies claim that the system will be able to identify suspicious persons and analyse their behavioural data to predict their future 'illegal' actions.⁴ As a result, many protesters and opposition leaders have been prosecuted, detained, and punished.⁵

Over the past decade, biometrics have been deployed to control mass protests on grounds of public security, restoring public order, and protecting third-parties' rights.⁶ Biometrics refer to computer-based systems able to perform remote,

¹ BBC, 'Russia's Use of Facial Recognition Challenged in Court' (*BBC News*, 31 January 2020) <www.bbc.com/news/technology-51324841>; Masha Borak, Inside Safe City, Moscow's AI Surveillance Dystopia (*Wired*, 6 February 2023) <<https://www.wired.com/story/moscow-safe-city-ntechlab/>>.

² Thomas Brewster, 'This Russian Facial Recognition Startup Plans To Take Its "Aggression Detection" Tech Global With \$15 Million Backing From Sovereign Wealth Funds' (*Forbes*, 22 September 2020) <www.forbes.com/sites/thomasbrewster/2020/09/22/this-russian-facial-recognition-startup-plans-to-take-its-aggression-detection-tech-global-with-15-million-backing-from-sovereign-wealth-funds/>.

³ Alessandro Mascellino, 'Rostec Turns Behaviour Analytics Development to "Smart" Anti-Riot Surveillance System' (*Biometric Update*, December 2021) <www.biometricupdate.com/202112/rostec-turns-behavior-analytics-development-to-smart-anti-riot-surveillance-system>.

⁴ ibid.

⁵ Anton Troianovski, Andrew E Kramer, and Andrew Higgins, 'In Aleksei Navalny Protests, Russia Faces Biggest Dissent in Years' (*The New York Times* (23 January 2021) <www.nytimes.com/2021/01/23/world/europe/navalny-protests-russia.html>.

⁶ Spiegel, 'G20 in Hamburg Wie die Behörden Extremisten vom Gipfel fernhalten wollen' (*Spiegel Panorama* (7 January 2017) <www.spiegel.de/panorama/gesellschaft/g20-extremisten-sollen-von-gipfel-in-hamburg-ferngehalten-werden-a-1155281.html>); José Ragas, 'La batalla por los rostros: el sistema de reconocimiento facial en el contexto del "estallido social" chileno' (2020) 14 Meridional Revista Chilena de Estudios Latinoamericanos 247; Paul Mozur, 'In Hong Kong Protests, Faces Become Weapons' (*The New York Times* (26 July 2019) <www.nytimes.com/2019/07/26/technology/hong-kong-protests-facial-recognition-surveillance.html>); Manish Singh, 'India Used Facial Recognition to Identify 1,100 Individuals at a Recent Riot' (*TechCrunch*, 11 March 2020) <<https://techcrunch.com/2020/03/11/india-used-facial-recognition-tech-to-identify-1100-individuals-at-a-recent-riot/>>.

real-time, and retrospective identification of persons based on the analysis of their physiological features.⁷ These systems have considerably extended the ability of government authorities to monitor, identify, and track individuals in public areas.⁸ Although they are marketed as being precise and objective,⁹ claims against their use allude to their indiscriminate, opaque, and biased nature.¹⁰ Furthermore, the lack of effective safeguards has led to their misuse and abuse.¹¹ While biometrics may support authorities in fulfilling their positive obligations regarding freedom of assembly, they may also contribute to shrinking this civic expression.¹²

Against that background, this chapter seeks to identify the legal implications of the use of biometrics on the right to freedom of assembly in the context of protests. Section 2 provides an overview of the use of biometrics to monitor assemblies, their promises, and justifications. Section 3 explores the broader socio-technical assemblage and practices that underpin their use by law enforcement, as well as their potential chilling effects on protesters and, consequently, the materialisation of assemblies. Considering its long-established relevant case law on the freedom to assembly generally, the chapter also discusses the approach of the European Court of Human Rights (ECtHR) to the nature and use of biometric data and surveillance technologies in the context of protests. Finally, section 5 offers a summary of relevant safeguards that may shield assemblies from the negative effects of biometrics.

2 Freedom of Assembly in the Age of Biometrics

The right to freedom of assembly allows people to come together and peacefully express ideas without unlawful interference. When such endeavour becomes violent, authorities are required to intervene to protect other relevant interests. In the quest to protect those interests, authorities have resorted to Artificial Intelligence (AI) technologies, with biometric systems emerging as the primary means to identify and track troublemakers.

⁷ Anil Jain, Ruud Bolle, and Sharath Pankanti, 'Introduction to Biometrics' in Anil Jain, Ruud Bolle, and Sharath Pankanti (eds), *Biometrics* (Springer 1996).

⁸ UN Human Rights Council, 'Report of the United Nations High Commissioner for Human Rights: Impact of New Technologies on the Promotion and Protection of Human Rights in the Context of Assemblies, including Peaceful Protests' (24 June 2020) UN Doc A/HRC/44/24, para 24.

⁹ Patrick J Grother, Mei L Ngan, and Kayee K Hanaoka, *Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification* (NIST 2018).

¹⁰ Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown 2016); Lotte Houwing, 'Opinions: Stop The Creep Of Biometric Surveillance Technology' (2020) 6 European Data Protection Law Review 174; Jenna Burrell, 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms' (2016) 1 Big Data & Society 3.

¹¹ Davide Castelvecchi, 'Is Facial Recognition Too Biased to Be Let Loose?' (2020) 587 Nature 347–49.

¹² UN Human Rights Council (n 8), para 24.

2.1 Freedom of Assembly: Positive Obligations and Biometric Promises

Demonstrations, protests, mass mobilisations, social movements, sit-ins, strikes, and rallies—both offline and online—are all forms of assembly widely protected by various international human rights instruments. On an international level, the right to freedom of assembly is protected by:

- Article 20 of the Universal Declaration of Human Rights (UDHR) and
- Article 21 of the International Covenant on Civil and Political Rights (ICCPR)

On a regional level, by:

- Article 11 of the African Charter on Human and Peoples' Rights (ACHPR),
- Article 15 of the American Convention on Human Rights (ACHR), and
- Article 11 of the European Convention on Human Rights (ECHR)

These instruments recognise the rights of individuals to peacefully gather with others in order to express and exchange common grievances and views.¹³ In that regard, governments are generally obliged not to interfere with peaceful assemblies—known as the negative obligation—and to actively facilitate the assembly and protect participants from external intervention—referred to as the positive obligation.¹⁴

That being said, judges have found that engaging in serious ‘reprehensible acts’ that put public order or third parties’ rights at risk may justify police intervention.¹⁵ After all, acts of violence—directly attributable to protesters—are not protected under the umbrella of freedom of assembly.¹⁶

Artificial intelligence (AI) technologies have brought novel dimensions to the preventative and investigative strategies of police and state security forces vis-à-vis protests. In their task of protecting other public interests and fulfilling their positive obligations, law enforcement has turned to biometrics to identify and investigate troublemakers. For instance, biometrics allowed law enforcement agencies to identify troublemakers during the G20 protests in Hamburg in 2017, the *Marcha del Millón* in Chile against social and economic inequality in 2019, the Hong Kong protests against the extradition law in 2019, and, more recently, during the 2021 protests in India against discriminatory legislation.¹⁷

¹³ International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR), art 21.

¹⁴ UNHRC ‘General Comment No 37, Article 21: Right of Peaceful Assembly’ (17 September 2020) UN Doc CCPR/C/GC/37 (General Comment No 37), para 23-4.

¹⁵ *Kudrevičius v Lithuania* App no 37553/05 (ECtHR, 15 October 2015), para 173-4.

¹⁶ *Gülçü v Turkey* App no 17526/10 (ECtHR, 6 June 2016), paras 110–7.

¹⁷ See n 6.

Biometric systems or ‘biometrics’¹⁸ are computer-based systems that link human physiological information to individuals’ identities.¹⁹ These systems are able to capture, convert, and compare various human features such as faces, voices, or fingerprints, with stored data of identified individuals.²⁰ To illustrate, the police may run face identification using a repository of images of suspected individuals, known as ‘persons of interest’. The system compares the processed data of any person that walks past the system’s camera against the images in the repository—referred to as biometric templates. When a match is found, it means that the scanned person is identified as the person of interest.

Despite the potential advantages that biometrics offer to law enforcement agencies, their use is done under the assumption that their functionality and effectiveness are indisputable truths; however, this is not the case.

2.2 Broken Promises

Biometrics have been advertised as being objective, having high accuracy rates, and outperforming humans in identification processes.²¹ However, biometrics are not neutral.²² By its very design, developers make deliberate choices by embedding and excluding certain options;²³ from deciding on the categories of data to be used to prioritising certain algorithmic techniques over others. Even the choice to not apply any sort of modelling is deliberate and has an impact on the system’s performance and decision-making process.²⁴ These choices often result in biased outputs that disproportionately impact individuals belonging to disadvantaged groups.²⁵ In the context of a protest, participants may be misrecognised due to a faulty algorithm, non-representative system’s trained data set, or mismatched target population and real-world deployment scenarios.²⁶

¹⁸ Catherine Jasserand, ‘Avoiding Terminological Confusion Between the Notions of “Biometrics” and “Biometric Data”: An Investigation into the Meanings of the Terms From a European Data Protection and a Scientific Perspective’ (2016) International Data Privacy Law 63.

¹⁹ International Organization for Standardization (ISO), ISO/IEC 2382-37:2012 ‘Information Technology—Vocabulary—Part 37: Biometrics’, Term 37.01.03, Note 4 <www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=55194>.

²⁰ Jain and others (n 7).

²¹ O’Neil (n 10).

²² ibid.

²³ Lucas Introna and David Wood, ‘Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems’ (2004) 2 Surveillance & Society 177, 179.

²⁴ Angelika Adensamer and Lukas Klausner, ‘Part Man, Part Machine, All Cop: Automation in Policing’ (2021) Frontiers in Artificial Intelligence 3.

²⁵ Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin’s Press 2018).

²⁶ Harini Suresh and John Guttag, ‘A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle’ in Equity and Access in Algorithms, Mechanisms, and Optimization (Association for Computing Machinery 2021).

In 2018, a study showed that FRT performed the poorest on dark-skinned females and the most accurately on white males with an error rate difference of 34 per cent.²⁷ These results were confirmed by the ‘Face Recognition Vendor Test on Demographic Effects’ conducted by the National Institute of Standards and Technology (NIST) in 2019.²⁸

Moreover, it is worth noting that while biometrics perform well and produce high accuracy rates when tested in controlled environments, they may well perform differently ‘in the wild’, particularly within large crowds such as during protests.²⁹ In spite of the low error rate advertised, a significant number of individuals may still be misidentified.³⁰ For instance, according to a report on FRT trials conducted by the London Metropolitan Police in 2018, out of forty-two matches produced by the system in 2018, only eight were accurate.³¹

In addition, designers have promoted ‘non-distinctive’ features called ‘soft biometrics’ which are often bundled with biometrics to enhance recognition performance. Soft biometrics refer to ancillary information gleaned from primary biometric data such as height, skin, hair, and eye colour.³² These systems are deemed valuable to control large crowds and have been used by the police, for instance, during the 2017 G20 protests in Hamburg. The biometric system ‘Videmo 360’ was advertised as being able to identify protesters but also detect, track, and categorise them according to their age, gender, and other behavioural data ‘even at low resolution and high-rotation’.³³ The details provided by these systems—when applied within protests—may reveal sensitive information such as individuals’ demographics, while also serving as indicators of crowd size but also emotion detection—violence—or group demographic composition.³⁴ This provides the police with actionable intelligence to determine what and who are the targets of interest.³⁵ For instance, if a hostile emotion or intention of violence is detected, this

²⁷ Joy Buolamwini and Timnit Gebru, ‘Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification’ (2018) Proceedings of Machine Learning Research 77–91.

²⁸ Patrick Grother, Mei Ngan, and Kayee Hanaoka, *Face Recognition Vendor Test (FYRT) Part 3: Demographic Effects* (NIST 2019).

²⁹ Nathalie Smuha and others, ‘How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission’s Proposal for an Artificial Intelligence Act’ (SSRN, 2021) 35 <<https://ssrn.com/abstract=3899991>>.

³⁰ FRA, ‘Facial Recognition Technology: Fundamental Rights Considerations in the Context of Law Enforcement’ (European Union Agency for Fundamental Rights, 21 November 2019) <<http://fra.europa.eu/en/publication/2019/facial-recognition-technology-fundamental-rights-considerations-context-law>>, 9.

³¹ Peter Fussey and Daragh Murray, ‘Independent Report on the London Metropolitan Police Service’s Trial of Live Facial Recognition Technology’ (2019) The Human Rights, Big Data and Technology Project 10.

³² See Antitza Dantcheva, Petros Elia, and Arun Ross, ‘What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics’ (2015) IEEE Transactions on Information Forensics and Security 11.

³³ Videmo, ‘Videmo 360’ <<http://videmo.de/en/products/videmo-sdk>>.

³⁴ Jungseock Joo and Zachary C Steinert-Threlkeld, ‘Image as Data: Automated Visual Content Analysis for Political Science’ (arXiv, 2 October 2018) <<http://arxiv.org/abs/1810.01544>>, 22.

³⁵ Fieke Jansen, Javier Sánchez-Monedero, and Lina Dencik, ‘Biometric Identity Systems in Law Enforcement and the Politics of (Voice) Recognition: The Case of SiiP’ (2021) Big Data & Society 4.

information may prompt police intervention during an assembly. However, claims that these systems can ‘reliably filter friends from foes, distinguish lies from truths, and use the tools of science to see inner worlds’³⁶ have been debunked.³⁷

Furthermore, the indiscriminate nature of biometrics is another cause for concern.³⁸ To identify persons of interest, the system has to compare the biometric data of every individual appearing in the image or footage frame against biometric templates. Consequently, innocent protesters and passers-by with no criminal record are subjected to biometric processing for no good reason.³⁹ This, for instance, happened in 2017 during the G20 protests in Hamburg where the biometric data of bystanders was collected and stored indefinitely.⁴⁰

Another critical aspect is the intrinsically opaque nature of the system processing.⁴¹ The most accurate biometrics are trained with complex algorithms that are not legible to humans.⁴² Protesters arrested immediately after a hit would not be able to challenge the decision if they do not know how—or why—the system reached that result.⁴³ Protesters cannot immediately challenge ‘smart’ system outputs, reducing their motivation to engage in protests.

3 Underlying Practices

The deployment of biometrics is determined by the assemblage of technical affordances and social practices.⁴⁴ Operators also play a significant role in how biometric performance and outputs are delivered. Police officers are usually the ones deciding where and when a biometric system will be deployed—organisational factors—what watch lists are chosen, how high will the threshold of similarity be—system factors—⁴⁵and how officers’ discretion and deference to systems outputs play a role in elucidating ways of suspicion ‘parameterised and primed’ through

³⁶ Kate Crawford, *Atlas of AI* (Yale UP 2022) 153.

³⁷ Arvind Narayanan, ‘AI Snake Oil, Pseudoscience and Hype’ in F Kaltheuner (ed), *Fake AI* (Meatspace Press 2021) 24.

³⁸ Houwing (n 10) 174.

³⁹ ibid.

⁴⁰ Der Hamburgische Beauftragte für Datenschutz und Informationsfreiheit, ‘Order Issued to Delete Biometric Database for Facial Matching in the Course of the G20 Investigation’ (2018) <<https://datenschutz-hamburg.de/pressemitteilungen/2018/12/2018-12-18-anordnung-biometrie>>.

⁴¹ See Jennifer Cobbe, ‘Administrative Law and the Machines of Government’ (2019) 39 *Legal Studies* 636, 639.

⁴² ibid.

⁴³ UN Human Rights Council (n 8) para 20.

⁴⁴ Mike Ananny and Kate Crawford, ‘Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability’ (2018) 20(3) *New Media & Society* 973–89; Kevin Haggerty and Richard Ericson, ‘The Surveillant Assemblage: Surveillance, Crime and Social Control’ (2000) *British Journal of Sociology* 605, 610.

⁴⁵ See Peter Fussey, Bethan Davies, and Martin Innes, ‘Assisted Facial Recognition and the Reinvention of Suspicion and Discretion in Digital Policing’ (2021) 61(1) *British Journal of Criminology* 325–44.

biometrics—operator factors.⁴⁶ In fact, police biases—selective adherence—tend to lean toward stopping and arresting individuals of certain groups more than others.⁴⁷ In fact, Black and other historically marginalised communities face higher risk of arrest, primarily because they are overrepresented in the ‘system’.⁴⁸ There is a ‘datafication’ of gang policing where young individuals from Black, Hispanic, and Arab groups constitute 90 per cent of gang databases, despite no real evidence linking them to actual gang involvement.⁴⁹

Another concerning issue in the use of biometrics is the lack of legal basis and function creep practices. Despite being granted permission by the Delhi High Court in 2018 to use FRT for the identification of missing children, the police went beyond and extensively used the system to identify ‘habitual protesters’ without judicial approval.⁵⁰ Moreover, it has also been observed that public authorities collect, retain, and process biometric data without a proper legal basis and mandatory safeguards. In 2020, the Dutch police with the help of FRT collected and stored more than 2.3 million photos of 1.3 million people, including people with no criminal background.⁵¹ Likewise, the Swedish, Italian, and Greek police were fined by their respective Data Protection Authorities (DPA) for employing biometric software without a proper legal basis and impact assessments.⁵²

The appeal to use biometrics not only relies on their claimed recognition accuracy⁵³ but also on their accessible infrastructure that made it possible for civilians and police officers to recognise the Capitol Hill rioters in January 2021 through web-based FRT platforms.⁵⁴ These platforms scrape the internet for images from social media and web profiles, comparing them to uploaded footage or images to identify and track individuals⁵⁵ from anywhere, at any time, by anyone.⁵⁶

⁴⁶ *ibid* 334.

⁴⁷ See Saar Alon-Barkat and Madalina Busuioc, ‘Human–AI Interactions in Public Sector Decision Making: “Automation Bias” and “Selective Adherence” to Algorithmic Advice’ (2023) 33 *Journal of Public Administration Research and Theory* 153.

⁴⁸ Karen Leavy, ‘Chilling Effects and Unequal Subjects: A Response to Jonathon Penney’s Understanding Chilling Effects’ (2022) 106 *Minnesota Law Review* 395.

⁴⁹ Alex Najibi, ‘Racial Discrimination in Face Recognition Technology, Harvard University, the Graduate School of Arts and Sciences’ (*SITN*, 24 October 2020) <<https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/>>.

⁵⁰ Singh (n 6).

⁵¹ Niels Waarlo and Laurens Verhagen, ‘De Stand van Gezichtsherkenning in Nederland’ *De Volkskrant* (27 March 2020) <www.volkskrant.nl/kijkverder/v/2020/de-stand-van-gezichtsherkenning-in-nederland~v91028/?referrer=https%3A%2F%2Fwww.google.com%2F>.

⁵² EDPB, ‘Swedish DPA: Police Unlawfully Used Facial Recognition App’ (EDPB, 12 February 2021) <https://edpb.europa.eu/news/national-news/2021/swedish-dpa-police-unlawfully-used-facial-recognition-app_en>.

⁵³ Grother, Ngan, and Hanaoka (n 9).

⁵⁴ See K Hill, ‘The Facial-Recognition App Clearview Sees a Spike in Use after Capitol Attack’ *The New York Times* (31 January 2021) <www.nytimes.com/2021/01/09/technology/facial-recognition-clearview-capitol.html>.

⁵⁵ *ibid*.

⁵⁶ To process biometric data, it suffices to have any repository of images. See Els Kindt, ‘Having Yes, Using No? About the New Legal Regime for Biometric Data’ (2018) 34(3) *Computer Law & Security Review* 1.

3.1 Chilling Effects

According to the UN Human Rights Council's General Comment on the Freedom of Peaceful Assembly, even though the collection of biometric data may assist authorities' obligations, such collection must not result in suppressing rights or creating a chilling effect on assembly participants.⁵⁷

It is important to note that the factors outlined above are cumulative and detrimental to freedom of assembly. Actual and future protesters will be deterred from participating in an assembly because they may be wrongfully identified by faulty biometrics, which is particularly problematic if they belong to vulnerable communities. Past experiences have shown the high likelihood of police abuse, the perception of biometric decisions as indisputable truths, and the absence of effective redress mechanisms. These factors create an overall hostile climate that influences participants' behaviour,⁵⁸ commonly referred to as 'chilling effects'.⁵⁹

In the words of Frederick Schauer, a chilling effect is an act of deterrence caused by the fear of retribution in the form of a fine, imprisonment, civil liability, or loss of government benefits.⁶⁰ In the context of a protest, the fear of being stopped and searched, arrested, or prosecuted for participating in a demonstration would discourage individuals from participating in it.⁶¹ According to Penney, there is also a social conformity effect in addition to the deterrent effect.⁶² The underlying idea here is that participants align their behaviour because of and according to anticipated potential sanctions.⁶³ The conforming behaviour here relates to the participant's compliance with social norms under the uncertainty of whether their conduct (protesting) is against the law.⁶⁴ In a manner, overt FRTs can give participants a sense of whether the protest is illegal or illegitimate. Many have argued, that sense of *uncertainty* here is key to enforcing disciplinary authority in a situation like a protest.⁶⁵ Individuals will make conscious decisions to self-restraint to avoid some perceived or explicit consequence of being identified, tracked, and

⁵⁷ General Comment No 37, para 61.

⁵⁸ Julian Staben, *Der Abschreckungseffekt auf die Grundrechtsausübung—Strukturen eines Verfassungsrechtlichen* (Mohr Siebeck 2016) 74.

⁵⁹ Valeria Aston, 'State Surveillance of Protest and the Rights to Privacy and Freedom of Assembly: A Comparison of Judicial and Protester Perspectives' (2017) 8 European Journal of Law and Technology 1.

⁶⁰ Frederick Schauer, 'Fear, Risk and the First Amendment: Unravelling the "Chilling Effect"' (1978) Boston University Law Review 689–90.

⁶¹ House of Commons, 'Joint Committee on Human Rights Legislative Scrutiny: Public Order Bill First Report of Session 2022' HC 351 HL Paper 16–23 [2022], 9.

⁶² Jon Penney, 'Understanding Chilling Effects' (2022) 106 Minnesota Law Review 1451; Aston (n 59) 2.

⁶³ ibid.

⁶⁴ ibid 1503.

⁶⁵ Marco Krüger, 'The Impact of Video Tracking Routines on Crowd Behaviour and Crowd Policing' in Lucas Melgaço and Jeffrey Monaghan (eds), *Protests in the Information Age: Social Movements, Digital Practices and Surveillance* (Routledge 2018) 146.

punished.⁶⁶ Those decisions will also have a supra-individual effect. Chilling effects limit the range of perspectives expressed in an assembly,⁶⁷ which is particularly relevant for its overall success.⁶⁸

In that regard, the UN High Commissioner stated that the use of biometrics to monitor protests jeopardises to some extent the ability of people to act and express themselves collectively.⁶⁹ Thus, authorities must justify any restriction, ensuring such measures are strictly necessary and proportionate to the permissible grounds.⁷⁰

4 Freedom of Assembly and the Use of Biometrics under the European Convention on Human Rights

Considering its long-established relevant case law on the freedom to assembly generally, this section discusses the approach of the ECtHR to the nature of biometric data and surveillance chilling effects within assemblies.

According to the ECtHR, Article 11 ECHR applies to peaceful assemblies and not to those with violent intentions from the outset or that reject the foundations of a democratic society.⁷¹ Assemblies can only be limited for the reasons provided by article 11(2) of the ECHR which include ‘the interests of national security or public safety, the prevention of disorder or crime, the protection of health or morals or the protection of the rights and freedoms of others’.⁷² Additionally, restrictive measures would require a clear and foreseeable legal basis, be *necessary in democratic society*—or have a pressing social need—, be proportionate⁷³, and accompanied by ‘relevant and sufficient’ reasons.⁷⁴

Catt and Segerstedt-Wiberg shed light on the sensitive nature of information revealing political opinion held by the police. In *Catt*, the ECtHR held that the applicants’ personal information, including a photograph of him taken at protests between 2005 and 2009, had a sensitive nature because they revealed their political opinions, thereby violating not only the right to privacy but also the right to

⁶⁶ Margot Kaminski and Shane Witnov, ‘The Conforming Effect: First Amendment Implications of Surveillance, Beyond Chilling Speech’ (2015) University of Richmond Law Review 483.

⁶⁷ Danielle Keats Citron and Daniel J Solove, ‘Privacy Harms’ (GWU Legal Studies Research Paper 2021) 54.

⁶⁸ Staben (n 58) 4.

⁶⁹ UN Human Rights Council (n 8).

⁷⁰ *ibid* para 36.

⁷¹ *Kudrevičius v Lithuania* (n 15), para 92.

⁷² Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR), art 11(2).

⁷³ Tor-Inge Harbo, *The Function of Proportionality Analysis in European Law* (Brill/Nijhoff 2015) 63.

⁷⁴ *Ezelin v France* App no 11800/85 (ECtHR, 26 April 1991), para 45; *Djavit An v Turkey* App no 20652/92 (ECtHR, 20 February 2003), para 63; *S and Marper v the United Kingdom* App nos 30562/04 and 30566/04 (ECtHR, 4 December 2008), para 101; *Kudrevičius v Lithuania* (n 15), paras 142–3.

freedom of assembly.⁷⁵ Although Mr Catt was never charged or accused with any illegal conduct during these protests, the police held this data in a database with the label ‘domestic extremism’.⁷⁶ The ECtHR stated that Mr Catt’s participation in the protest was a democratic act protected by article 11 of the ECHR and held that the circumstances of the retention must have had a chilling effect on Mr Catt’s legitimate right to engage in a public protest.⁷⁷ It also sets precedent to future participants, discouraging them from taking part in similar gatherings.⁷⁸

Although there is no case before the ECtHR involving the retention of biometric data in the context of a protest, the ECtHR has generally a similar view on the sensitive nature of biometric data such as DNA and fingerprints for having the potential to reveal health and demographic information, especially when processed through automated means.⁷⁹ The ECtHR has also considered the ability to subject mere custody photographs (raw biometric data) held by the police to automated facial recognition and facial mapping techniques.⁸⁰

Moreover, although not discussed further in *Catt*, the lack of a ‘concrete suspicion’ or ‘serious threat’ that justified the retention of his personal data is particularly problematic. This is because Mr. Catt had no criminal record and his conduct was demonstrably peaceful. Thus, the indefinite retention of his information was not relevant to police operations.⁸¹ In previous decisions, such as *Schwabe*, the offence that the applicant was feared to commit, leading to his detention, was not ‘sufficiently concrete and specific’,⁸² nor had the government proved that the threat of committing such an offence was imminent.⁸³ It follows that the collection of a protester’s personal data without regard to those elements put protesters who are entitled to the presumption of innocence in the same place as those who have been convicted.⁸⁴ As mentioned in section 2.2, these practices are currently the rule in many police operations.⁸⁵

This point also raises the issue of the indiscriminate nature of biometrics, which is analogous to the use of general, indiscriminate, or non-targeted surveillance measures by the government. Against all precedents, in *Big Brother Watch*,

⁷⁵ *Catt v the United Kingdom* App no 43514/15 (ECtHR, 24 January 2019), para 112; *Segerstedt-Wiberg v Sweden* App no 62332/00 (ECtHR, 6 September 2006), para 107.

⁷⁶ *Catt v the United Kingdom* (n 75), para 123.

⁷⁷ *ibid*.

⁷⁸ *Zakharov and Varzhabetyan v Russia* App nos 35880/14 and 75926/17 (ECtHR, 13 January 2021), para 90

⁷⁹ *S and Marper v the United Kingdom* (n 79), paras 72–78; *PN v Germany* App no 74440/17 (ECtHR, 16 November 2020), para 84.

⁸⁰ *Gaughran v United Kingdom* App no 45245/15 (ECtHR, 13 June 2020), para 70. See Kindt (n 56); Philipp Terhörst and others, ‘On Soft-Biometric Information Stored in Biometric Face Embeddings’ (2021) 4 IEEE Transactions on Biometrics, Behavior, and Identity Science 519.

⁸¹ *Catt v the United Kingdom* (n 75) para 23.

⁸² *Schwabe and MG v Germany* App nos 8080/08 and 8577/08 (ECtHR, 1 March 2012), para 82.

⁸³ *ibid* para 85.

⁸⁴ Jake Goldenfein, *Monitoring Laws: Profiling and Identity in the World State* (CUP 2019) 57.

⁸⁵ HmbBfDI (n 40).

the ECtHR held that the bulk interception of communications by government agencies can be compatible with the ECHR.⁸⁶ In the ECtHR's view, non-targeted surveillance is not incompatible with the ECHR, but the lack of safeguards that prevent unfettered discretion and disproportionate restrictions.⁸⁷ Among those safeguards are the statement of clear and specific rules, storage periods, authorisation, and independent oversight.⁸⁸ The problem with this approach is that all the burden is placed on 'effective' safeguards without questioning the merits of the measures in the first place as generalised and non-targeted surveillance is categorically disproportionate.⁸⁹ In this regard, Judge Pinto de Albuquerque pointed out, in his dissenting opinion, the ECtHR's bias and lack of technical knowledge when assessing the actual effectiveness of bulk surveillance measures.⁹⁰ This is especially relevant for the use of biometrics in public spaces since claims over their effectiveness and necessity are still an issue.

The use of biometrics in a protest raises doubts about their compatibility with the ECHR as broad and indiscriminate measures against an abstract—non-concrete or specific—threat may not tilt the balance,⁹¹ instead could result in government abuse.⁹²

5 Effective Safeguards in Place?

Allowing the use of live biometrics may only be justified under specific grounds, that is, as national security or public order, provided they are effective to achieve those and necessary and proportionate to the seriousness of the threat. Thus, there must be sufficient reasons to think that an imminent and serious crime is likely to occur. In those cases, the establishment of safeguards is however essential to mitigate the likelihood of harms posed on freedom of assembly.

According to the United Nations High Commissioner for Human Rights, authorities should refrain from collecting biometric data on peaceful protesters to harass, intimidate, track, or stigmatise them.⁹³

⁸⁶ *Big Brother Watch v the United Kingdom* App nos 58170/13, 62322/14, and 24960/15 (ECtHR, 25 May 2021), para 287; *Weber and Saravia v Germany* App no 54934/00 (ECtHR, 29 June 2006), paras 123, 125; *Szabo v Hungary* App no 37138/14 (ECtHR, 12 January 2016), para 73.

⁸⁷ *Cumhuriyet Vakfi and Others v Turkey* App no 28255/07 (ECtHR, 8 October 2013), para 63.

⁸⁸ *Szabo v Hungary* (n 86).

⁸⁹ Marko Milanovic, 'The Grand Normalisation of Mass Surveillance: ECtHR Grand Chamber Judgments in Big Brother Watch and Centrum för Rättvisa' (*EJIL: Talk*, 26 May 2021) <www.ejiltalk.org/the-grand-normalization-of-mass-surveillance-echr-grand-chamber-judgments-in-big-brother-watch-and-centrum-for-rattvsa/>.

⁹⁰ Judge Albuquerque, Partly Dissenting Opinion in *Big Brother Watch v the United Kingdom* (n 86) para 8.

⁹¹ Ilia Siatitsa, 'Freedom of Assembly Under Attack: General and Indiscriminate Surveillance and Interference with Internet Communications' (2020) International Review of the Red Cross 191.

⁹² *ibid.*

⁹³ UN Human Rights Council (n 8) para 24.

While the protection of protesters biometric data by data protection and privacy legislation is vital, it is important to recognise their broader human rights implications. In that regard, Scassa suggests that keeping control over personal data should be done with a the goal of fostering the exercise of a variety of human rights, including freedom of assembly.⁹⁴

Thus, technologies must be developed with the safeguarding of human rights as the pre-eminent goal.⁹⁵ To that end, some scholars propose a governing human rights approach to AI technologies that include:

- an integrated set of technical and organisational mechanisms;
- an independent external oversight; and
- channels for public deliberation from the start of the technology development.⁹⁶

Among technical safeguards are: ensuring that systems are robust and accurate; have data quality; transparency; and human oversight. Many data protection laws impose safeguards against security flaws, function creep, and privacy risks. For example, if biometrics are stored in a centralised database, security requirements must then be strengthened by using, for example, encryption or shifting to localised databases.⁹⁷ Also, the data fed into the system must comply with all data quality requirements, including data accuracy, relevancy, and diversity to prevent failure repetitions but also to mitigate biased decisions as a result of having poor quality input.⁹⁸ Organisational measures may involve conducting risk impact assessments, implementing procedures to limit storage periods, establishing fall back procedures, and implementing appeal and redress mechanisms.⁹⁹

Oversight mechanisms and authorisation are particularly crucial in the assessment of legal compliance beyond data protection and privacy safeguards.¹⁰⁰ As stated by the Human Rights Committee, independent and transparent scrutiny and oversight must be exercised in the collection and use of personal information

⁹⁴ Teresa Scassa, 'A Human Rights-Based Approach to Data Protection in Canada' in Elizabeth Dubois and Florian Martin-Bariteau (eds), *Citizenship in a Connected Canada: A Research and Policy Agenda* (University of Ottawa Press 2020) 183.

⁹⁵ Alessandro Mantelero, 'Report on Artificial Intelligence and Data Protection: Challenges and Possible Remedies' (Council of Europe 2019) 6.

⁹⁶ Karen Yeung, Andrew Howes, and Ganna Pogrebna, 'AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing' in Markus D Dubber and others (eds), *Oxford Handbook of AI Ethics* (OUP 2019) 87.

⁹⁷ European Data Protection Board, 'Guidelines 3/2019 on Processing of Personal Data through Video Devices Version 2.0' (29 January 2020), para 88.

⁹⁸ See Minister's Deputies (Council of Europe) Recommendation CM/Rec (2020)1 of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems, para 4.4.

⁹⁹ Els Kindt, *Privacy and Data Protection Issues of Biometric Applications* (Springer 2016) 881.

¹⁰⁰ See Thorsten Wetzling and Kilian Vieth, 'Legal Safeguards and Oversight Innovations for Bulk Surveillance: An International Comparative Analysis' in Lora Anne Viola and Paweł Laidler (eds), *Trust and Transparency in an Age of Surveillance* (Routledge 2021) 151.

and data regarding peaceful assemblies.¹⁰¹ Usually judges are called upon for this task due to their qualifications, experience, and expertise in applying the law. Combining their abilities with experts' technical assessment is essential to understand the human rights implications of a given situation.¹⁰² That being said, the Special Rapporteur on the rights to freedom of peaceful assembly and of association held that non-judicial oversight should also be implemented.¹⁰³ This includes the establishment of an independent oversight bodies with full investigative powers to handle complaints objectively, fairly, and promptly.¹⁰⁴ But also enabling the direct participation of affected stakeholders, including citizens, through participatory co-design models.¹⁰⁵

Although safeguards cannot absolutely prevent authorities' abuse, they lessen its likelihood of occurrence. Developing safeguards aligned with human rights for AI is necessary not only to ensure legal compliance but to strengthen trust in the responsible adoption of these technologies.

6 Conclusion

Biometric systems may assist law enforcement agencies in fulfilling their positive obligations concerning the right to assembly. However, the opaque, indiscriminate, and error-prone nature of biometric systems, together with the disproportionate police practices behind their use can have the potential to disrupt and deter peaceful assemblies. Under the view of the ECtHR, the use of general and indiscriminate surveillance technologies may be permissible under certain conditions. The Court has emphasised the need for effective safeguards to balance the adverse effects of such surveillance. While recognising that safeguards alone are not sufficient to justify restrictive measures but the actual effectiveness and necessity of such use, they play a critical role in mitigating risks to human rights, including the right to freedom of assembly.

¹⁰¹ General Comment No 37, para 63.

¹⁰² Murray Daragh and others, 'Effective Oversight of Large-Scale Surveillance Activities: A Human Rights Perspective' (2020) 11 *Journal of National Security Law and Policy* 749.

¹⁰³ Human Rights Council, 'Joint report of the Special Rapporteur on the rights to freedom of peaceful assembly and of association and the Special Rapporteur on extrajudicial, summary or arbitrary executions on the proper management of assemblies' (UN Doc A/HRC/31/66, 4 February 2016), para 96(b).

¹⁰⁴ *ibid* para 96(d).

¹⁰⁵ Jesper Simonsen and Toni Robertson (eds), *Routledge International Handbook of Participatory Design* (Routledge 2013).

Artificial Intelligence and the Right to Property

The Human Rights Dimension of Intellectual Property

*Letizia Tomada and Raphaële Xenidis**

1 Introduction

In the era of fast-developing artificial intelligence (AI) technologies, the tension between proprietary rights and transparency requirements represents a key regulatory issue in the context of fundamental rights. The legislator has to some extent acknowledged the need of achieving a balanced interplay between safeguarding intellectual property rights (IPRs) and ensuring transparency in intellectual property (IP) law and data protection law (eg in the General Data Protection Regulation (GDPR) of the European Union (EU)). Yet, the currently applicable legal framework presents challenges of implementation and limitations that deserve particular attention. To this end, stemming from the debate on the human rights dimension of IP, the present contribution first examines to what extent certain AI applications can be IP protected (section 2). Second, it explores the consequences on the human right to property and the delicate interplay with other fundamental rights (section 3). Further, it analyses IP law exceptions as regulatory solutions aimed at overcoming potential human rights clashes (section 4) and proposes ways forward (section 5) to achieve more balance between the rights involved.

2 The Implementation of the Right to Property: Relevance for Certain AI Applications

This section explores the interplay between IP and human rights in order to lay the foundations for investigating the implications of AI on the right to property, framed as IP. To this aim, the chapter first considers how this relationship is dealt with in relevant international treaties and presents the debate over the recognition

* This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 898937.

of ‘human rights status’ to IP. Second, bringing AI into the equation, it analyses to what extent AI systems can be subject to IP protection.

2.1 Intellectual Property and Human Rights in International Treaties

Early international human rights instruments already showed a certain degree of attention for IP. For instance, the American Declaration on the Rights and Duties of Man granted protection to the interests of individuals in their literary, scientific, or artistic works—including their inventions—within the broader right to enjoy the benefits of science and culture.¹ Article 15(1)(c) of the International Covenant on Economic, Social and Cultural Rights (ICESCR), adopted in 1966, similarly protects everyone’s right to ‘benefit from the protection of moral and material interests resulting from any scientific, literary or artistic production of which he is the author’.² These instruments have widely influenced regional treaties, declarations, and national constitutions.³ In Europe, article 17(2) of the 2000 EU Charter of Fundamental Rights (EU Charter) mentions the protection of IP explicitly.⁴ By contrast, article 27 of the Universal Declaration of Human Rights (UDHR), article 1 of the Additional Protocol (A1P1) to the European Convention on Human Rights (ECHR), and article 14 of the African Charter on Human and Peoples’ Rights (ACHPR)—all protecting the right to property—do not mention its intellectual dimension.⁵

Conversely, the first international IP law treaties did not include human rights considerations. Indeed, as Helfer and Austin note,⁶ both the Paris and Berne Conventions were signed at the end of the nineteenth century and, therefore, before the establishment of the international human rights framework. The approach partly changed with the Agreement on the Trade-Related Aspects of Intellectual Property Rights (TRIPS Agreement), signed in 1994, which, despite not recognising human rights as such, includes provisions concerning the implications of IP protection on human rights. In particular, article 7 of the TRIPS Agreement

¹ American Declaration on the Rights and Duties of Man (1948), art 13.

² International Covenant on Economic, Social and Cultural Rights (ICESCR), art 15(1).

³ For more details in this regard and examples, see G Spina Ali, ‘Intellectual Property and Human Rights: A Taxonomy of Their Interactions’ (2020) 51 International Review of Intellectual Property and Competition Law 411, 413.

⁴ EU Charter, art 17.

⁵ Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR), art 27; Additional Protocol to the European Convention on Human Rights, art 1 (A1P1); African Charter on Human and Peoples’ Rights (1981), art 14. That said, the right to property still applies to intellectual property, see eg *Anheuser-Busch Inc v Portugal* App no 73049/01 (ECtHR, 11 January 2007), para 72.

⁶ L Helfer and G Austin, *Human Rights and Intellectual Property: Mapping the Global Interface* (CUP 2011); see also Spina Ali (n 3) 414.

provides that IP protection needs to be balanced with other rights and obligations to safeguard social and economic welfare. In addition, article 8 allows contracting parties to implement measures necessary to protect public health and to promote the public interest in sectors of vital importance.⁷ In light of this background, section 2.2 explores to what extent IPRs have been considered as human rights.

2.2 Intellectual Property as a Human Right

Intellectual property rights may be considered different from other human rights such as the freedom of expression, religion, or association because their protection comes with a series of limitations linked to regulation in the name of the general or public interest.⁸ Nevertheless, the law recognises the human rights dimension of IP either by (i) elevating IPRs to the same level as human rights and establishing a right to IP or (ii) extending the safeguards linked to tangible property to IPRs.⁹

First, the idea of equating IPRs to human rights has triggered several critiques.¹⁰ It has been argued that this approach would excessively reinforce IP claims over countervailing rights, such as freedom of expression or information, which should have priority.¹¹ Furthermore, the recognition of IPRs as a right *per se* rather than as a temporary monopoly granted for the benefit of the community, is deemed to mark a departure from their public-interest underpinnings.¹² Currently, no international convention or regulatory instrument regards IPRs as being at the same level as human rights, but some recognise the human rights dimension of certain features of the IP legal framework. For example, article 15 of the ICESCR acknowledges ‘the right of everyone to benefit from the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author’.¹³ Although the scope of the provision does not cover trademarks or trade secrets,¹⁴ and it is controversial whether the right can also be recognised for scientific inventions,¹⁵ it is uncontested that it does encompass the copyright

⁷ TRIPs Agreement, art 8(1).

⁸ See eg EU Charter, art 17(1), which recognises exceptions ‘in the public interest and in the cases and under the conditions provided for by law’ and indicates that ‘[t]he use of property may be regulated by law in so far as is necessary for the general interest’.

⁹ Spina Alí (n 3) 418.

¹⁰ *ibid.*

¹¹ R Ostergaard, ‘Intellectual Property: A Universal Human Right?’ (1999) 21 *Human Rights Quarterly* 156.

¹² L O’Mellin, ‘Software and Shovels: How the Intellectual Property Revolution is Undermining Traditional Concepts of Property’ (2007) 76 *University of Cincinnati Law Review* 143, 144.

¹³ ICESCR, art 15(1); see also UDHR, art 27.

¹⁴ M Carpenter, ‘Intellectual Property: A Human (Non-Corporate) Right’ in D Keane and Y McDermott (eds), *The Challenge of Human Rights: Past, Present and Future* (Edward Elgar 2012) 322.

¹⁵ Rochelle Cooper Dreyfuss, ‘Patents and Human Rights: Where is the Paradox?’ (New York University Law School, Public Law Research Paper No 06-29, 2006) <<https://ssrn.com/abstract=929498>>.

of authors. In this context, the constitutions of certain countries go a step further thereby establishing a ‘right to intellectual property’ and thus ensuring a proprietary right to artists and innovators over their endeavours,¹⁶ in some cases restricting it only to some forms of protection, copyright in particular.¹⁷

Second, the recognition of the human rights status of IPRs can derive from an extension of the protection of the right to property. Article 1 of Protocol 1 (A1P1) to the ECHR establishes the right to ‘the peaceful enjoyment of [one’s] possessions’ and the right not to be unlawfully deprived of them. There are different scholarly views on whether IP should be equated to tangible property. On the one hand, it has been suggested that IP should not be assimilated to the traditional form of property, since it lacks the features of exclusivity and rivalry which characterise the latter.¹⁸ On the other hand, others argue that both forms of property share the same justifications, based on the principle that natural persons shall have the right to enjoy the fruits of their labour.¹⁹ In any case, it is relevant to note that, pursuant to the European Court of Human Rights’ (ECtHR) interpretation, the concept of ‘possession’ enshrined in A1P1 ECHR is not limited to physical goods and does not depend on national law classifications.²⁰ The concept can therefore be stretched so as to include ‘second generation’ human rights (in particular the economic dimension) and IPRs.²¹ As a result, both the national and international jurisprudence very often assimilate IP to tangible property, thereby extending to the former the provisions that apply to the latter. The European Commission on Human Rights (ECommHR) first set this approach in *Smith Kline v the Netherlands*. Even if the Commission dismissed the case on the ground that deciding to grant a compulsory licence for a patented compound falls within a government’s margin of appreciation, it clearly held that a patent shall be considered a possession.²² The ECommHR further reaffirmed that A1P1 encompasses IPRs in cases relating to

¹⁶ Eg Constitution of Colombia (1991), art 61; Constitution of Kenya (2010), art 69; Constitution of Montenegro (2007), art 76; EU Charter on Fundamental Freedoms (2000), art 17(2).

¹⁷ Eg Constitution of Sweden (1974), art 16; Constitution of Albania (2008), art 58; Constitution of Portugal (1976), art 42(2). For additional examples and further analysis in this regard, see Spina Ali (n 3) 420.

¹⁸ V Heins, ‘Human Rights, Intellectual Property and Struggles for Recognition’ (2007) 9 *Human Rights Review* 213, 217.

¹⁹ C Geiger, ‘Reconceptualizing the Constitutional Dimension of Intellectual Property’ in P Torremans (ed), *Intellectual Property and Human Rights* (Wolters Kluwer 2015) 115, 118; C Geiger, ‘Implementing Intellectual Property Provisions in Human Rights Instruments: Towards a New Social Contract for the Protection of Intangibles’ in C Geiger (ed), *Research Handbook on Human Rights and Intellectual Property* (Edward Elgar 2015) 661.

²⁰ See eg *Beyeler v Italy* App no 33202/96 (ECtHR, 5 January 2000), para 100; *Matos and Silva Lda v Portugal* App no 15777/89 (ECtHR, 16 September 1996), para 75; *Former King of Greece v Greece* App no 25701/94 (ECtHR, 28 November 2002), para 60; *Forrer-Niedenthal v Germany* App no 47316/99 (ECtHR, 20 February 2003), para 32; and *Broniowski v Poland* App no 31443/96 (ECtHR, 28 September 2005), para 129.

²¹ ECtHR, ‘Guide on Article 1 of Protocol No 1: Protection of Property’ (31 August 2021) 1, 14 <www.echr.coe.int/Documents/Guide_Art_1_Protocol_1_ENG.pdf>.

²² *Smith Kline and French Laboratories Ltd v the Netherlands* App no 12633/87 (ECommHR, 4 October 1990).

patents and copyright.²³ The Court further corroborated this principle in several rulings.²⁴ For example, in *Dima v Romania* the ECtHR affirmed that copyright is a form of property that comes in existence at the moment of the creation of the work,²⁵ thereby assimilating copyright to the right to property and extending to the former the safeguards and provisions that apply to the latter. In any case, despite the different approaches on the matter, it is clear that the discussions on the human rights dimension of IP only come into consideration when the subject matter can be IP protected.

2.3 IP Protection of AI Applications

In order to assess the implications of AI on the right to property, framed as a right to IP, it is first necessary to investigate to what extent AI applications can be IP protected. Indeed, AI systems can be covered by IP protection via patents, copyright, and trade secrets. The present sub-section 2.3 briefly explores the different means of AI IP protection and analyses their respective uses and related challenges. While international law instruments, such as the TRIPs Agreement and the Paris and Berne Conventions set common standards for IP protection, for the purposes of the present contribution it is necessary to focus on their implementation at regional or national level. To this aim, Europe is considered as an example. Moreover, the discussion is limited to the protection of the AI systems *as such* and is therefore not extended to the AI output.

2.3.1 Patent Protection

Despite the proliferating number of patent applications in the field of AI,²⁶ the patent route for AI protection poses several legal issues and uncertainties.²⁷ The

²³ See eg *Lenzing AG v United Kingdom* App no 38817/97 (ECommHR, 9 September 1998); and *Aral and Tekin v Turkey* App no 24563/94 (ECommHR, 14 January 1998), para 4.

²⁴ For an overview of the relevant rulings to date, see ECtHR (n 21); for an analysis of some of the cases, see Spina Alí (n 3) 422–24.

²⁵ *Dima v Romania* App no 58472/00 (ECtHR, 26 May 2005), paras 8–26. The case in particular concerned the design of the new national emblem of the Romanian state commissioned by the Romanian parliament to the applicant, who had not been awarded compensation and had seen his claims dismissed by national courts on the grounds that copyright does not cover state emblems, which fall in the public domain. Despite assimilating copyright to the right to property, in the case at hand, the ECtHR decided in favour of the state, considering that it is for national courts to decide on the extent and existence of property rights.

²⁶ An EPO study highlights that since 2011 inventions in the AI-field presented an average annual growth rate of 43 per cent with 83 patent applications in 2016. See European Patent Office (EPO), ‘Patents and the Fourth Industrial Revolution: The Inventions Behind Digital Transformation’ (December 2017) <[http://documents.epo.org/projects/babylon/eponet.nsf/0/17FDB5538E87B4B9C12581EF0045762F/\\$File/fourth_industrial_revolution_2017_en.pdf](http://documents.epo.org/projects/babylon/eponet.nsf/0/17FDB5538E87B4B9C12581EF0045762F/$File/fourth_industrial_revolution_2017_en.pdf)>.

²⁷ These include not only the subject matter eligibility, but also the requirement of sufficiency of disclosure and the balance between promotion of or hindrance to innovation. See M Iglesias and others, *Intellectual Property and Artificial Intelligence: A Literature Review* (2019) (Publications Office of the European Union 2021) 1, 6–7.

question of whether AI may be eligible for patent protection is inextricably linked to the debate concerning the patentability of computer software. Pursuant to the European Patent Convention (EPC), computer programs as such are excluded from patent protection.²⁸ Yet, computer programs that have a technical character—and thus produce a further technical effect when running on a computer—can be patented. Similarly, mathematical methods as such are excluded from patentability,²⁹ but the exclusion applies only if the claim concerns a purely abstract mathematical method and does not relate to any technical means.³⁰ Since AI and machine learning (ML) inventions are based on algorithms and computational models, which are ‘per se’ of an abstract mathematical nature, irrespective of whether they can be “trained” on training data,³¹ the guidance on computer programs and mathematical methods can also be used for AI and ML related inventions.³² Therefore, they are not excluded from patentability as long as the subject matter presents a technical character; that is, if the claim is directed to a method involving the use of technical means, such as a computer.³³ For example, classifying abstract data records without any indication of the technical use made of the resulting classification is not per se a technical purpose,³⁴ while using a neural network in a health-monitoring system for identifying irregular heartbeats represents a technical contribution instead.³⁵

2.3.2 Copyright Protection

Furthermore, when it comes to copyright protection, the Software Directive provides that computer programs are protected, but protection does not extend to ‘ideas, procedure, methods of operation or mathematical concepts’.³⁶ In particular, copyright protection applies only to the expression of the computer program and does not cover the ideas, logic, algorithms, and underlying programming languages instead.³⁷ Therefore, copyright protects the source code and object code implementing the algorithm,³⁸ while the program’s functionality, data formats, and programming languages are interpreted as ‘ideas’ and are therefore not covered.³⁹

²⁸ EPC, art 52(2)(c), (3).

²⁹ EPC, art 52(2)(a), (3).

³⁰ EPO ‘Guidelines for Examination G-II 3.3, Artificial intelligence and machine learning’, European Patent Office (EPO), 2019 update.

³¹ *ibid* G-II 3.3.1.

³² *ibid*.

³³ *ibid*.

³⁴ EPO T 1784/06 (Classification Method/COMPTEL) ECLI:EP:BA:2012:T178406.20120921 point 3.1.1. (21 September 2012).

³⁵ EPO Guidelines for Examination G-II 3.3.1.

³⁶ Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs [2009] OJ L111/16 (Software Directive), art 1(2).

³⁷ *ibid* Recital 11 which reads: ‘to the extent that logic, algorithms and programming languages comprise ideas and principles, those ideas and principles are not protected’.

³⁸ Case C-393/09 *Bezpecnostní Softwarová Asociace* EU:C:2010:816, paras 28–42.

³⁹ Case C-406 SAS Institute EU:C:2012:259, paras 38–46.

Nevertheless, algorithms can still be eligible for copyright protection and constitute a ‘work’ under the InfoSoc Directive⁴⁰ to the extent that they fulfil two criteria. First, they must satisfy the criteria of originality, by being the author’s own original intellectual creation. Second, the ‘work’ shall be an expression of such a creation and not be dictated by technical functionality.⁴¹ Given that, for several types of algorithmic models, it may well be challenging to meet these criteria, businesses often rely on trade secrets protection.

2.3.3 Trade Secrets Protection

The EU Trade Secrets Directive (EUTSD) harmonises trade secrets protection in the EU. It covers know-how, technological and business information, and commercial data.⁴² Although trade secrets do not ‘create an exclusive right to know-how or information’⁴³ and they did not originate as an instrument of IP protection but rather as a safeguard of the confidentiality of commercial information, developers often rely on trade secrets law to protect their algorithms.⁴⁴ In fact, AI algorithms are often neither generally known nor easily accessible to the public, they have commercial value and are usually kept secret, thereby meeting the necessary requirements for trade secrets protection.⁴⁵

3 Recognising a Fundamental Right to Intellectual Property in the Context of AI Applications: What Consequences?

Recognising a human right dimension to IP in the context of AI systems would create complex interactions with other provisions of the human rights catalogue. On the one hand, by protecting innovation and encouraging investment, it would arguably entail positive consequences for the freedom to conduct a business anchored, for example, in article 16 of the EU Charter. A human right to IP in the AI context could also foster the enjoyment of related rights such as the freedom

⁴⁰ Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society [2001] OJ L167/10–19 (InfoSoc Directive), arts 2–4.

⁴¹ See eg Joined Cases C-403/08 and C-429/08 *Football Association Premier League and Others* EU:C:2011:631, para 97; Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* EU:C:2009:465, para 39. For an analysis on which types of algorithmic models may or may not satisfy the relevant criteria, see K Foss-Solbrekk, ‘Three Routes to Protecting AI Systems and Their Algorithms under IP Law: The Good, the Bad and the Ugly’ (2021) 16 Journal of Intellectual Property Law and Practice 247, 251.

⁴² Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure [2016] OJ L157/1 (Trade Secrets Directive/EUTSD), Recitals 1, 2, 14.

⁴³ *ibid*, Recital 16.

⁴⁴ Guido Noto La Diega, ‘Against the Dehumanisation of Decision-Making’ (2018) 9(3) Journal of Intellectual Property, Information Technology and Electronic Commerce Law 12..

⁴⁵ *ibid* art 2(1)(a)–(c).

of (commercial) expression,⁴⁶ as well as the right to pursue a chosen economic activity included in the right to work in article 1 of the European Social Charter, article 6(1) of the ICESCR, and article 15 of the EU Charter.⁴⁷

On the other hand, applying the human right to property to AI systems could create important clashes with other human rights either directly or because it could threaten their enforcement. In its 2020 resolution on 'Intellectual property rights for the development of artificial intelligence technologies', the European Parliament 'stresse[d] that the protection of intellectual property must always be reconciled with other fundamental rights and freedoms'.⁴⁸ Recent debates over the fundamental rights risks associated with the increasing use of AI applications have highlighted the importance of two particular aspects: transparency and explainability. The lack of transparency and explainability of AI systems is liable to (i) directly infringe on the freedom to seek, receive, and give information related to the freedom of expression; and (ii) indirectly infringe on other human rights by rendering more difficult or altogether preventing their exercise and enforcement.

At the level of direct clashes, IPRs protecting AI systems could negatively impact on rights related to the freedom of expression.⁴⁹ These are anchored in article 11 of the EU Charter and article 10 of the ECHR, and include the freedom to receive and distribute information and the freedom to hold and share opinions. Intellectual property rights could prevent end users of AI systems from obtaining information on the logic, rationale, and circumstances of the decision-making process involved. For example, in the context of algorithmic content moderation, the IP protection of the AI systems used could clash with the freedom of expression of online platforms and social media users.⁵⁰ The non-disclosure of the parameters and the functioning of the algorithmic decision-making system used could amount to an impossibility to ensure that certain (categories of) users and content are not systematically silenced or censored by automatic content moderation. The right to consumer protection⁵¹ might also be affected negatively, for example,

⁴⁶ Recognised, for example, in *Krone Verlag GmbH & Co KG v Austria* (No 3) App no 39069/97 (ECtHR, 11 December 2003). See European Union Agency for Fundamental Rights, 'Freedom to Conduct a Business: Exploring the Dimensions of a Fundamental Right' (2015), 10 <https://fra.europa.eu/sites/default/files/fra_uploads/fra-2015-freedom-conduct-business_en.pdf>.

⁴⁷ EU Agency for Fundamental Rights (n 46).

⁴⁸ European Parliament, 'Resolution of 20 October 2020 on Intellectual Property Rights for the Development of Artificial Intelligence Technologies 2020/2015(INI)' (2020) <www.europarl.europa.eu/doceo/document/TA-9-2020-0277_EN.html>.

⁴⁹ Rikke Frank Jørgensen, 'The Right to Express Oneself and to Seek Information' in Rikke Frank Jørgensen, Ernest J Wilson, and William J Drake (eds), *Human Rights in the Global Information Society* (MIT Press 2006) 55; Robin Gross, 'Information Property Rights and The Information Commons' in Rikke Frank Jørgensen, Ernest J Wilson, and William J Drake (eds), *Human Rights in the Global Information Society* (MIT Press 2006) 109.

⁵⁰ See the chapter by Giovanni De Gregorio and Pietro Dunn in this volume.

⁵¹ For a broader discussion of consumer protection as a human right, see Sinai Deutch, 'Are Consumer Rights Human Rights?' (1994) 32(3) Osgoode Hall Law Journal 537.

when IPRs prevent users from availing of the interoperability of different AI systems.⁵²

The proprietary protection of AI systems can also create a number of indirect clashes with other human rights, namely by rendering more difficult or altogether preventing their exercise and enforcement. Transparency and explainability are often framed as *sine qua non* conditions for enforcing a series of fundamental rights such as the right to non-discrimination (article 14 of the ECHR and article 21 of the EU Charter), the right to privacy or respect for private and family life (article 8 of the ECHR and article 7 of the EU Charter), the right to protection of personal data (article 8 of the EU Charter) the right to a fair trial and to an effective remedy (article 6 of the ECHR and article 47 of the EU Charter), and so on. In many cases, transparency and explainability condition the enforcement of those rights because they affect victims' ability to bring evidence of breaches as well as judges' ability to assess the merits of those cases. Framing IPRs as fundamental rights might trigger protective safeguards that clash with such transparency and explainability requirements. This could, in turn, make victims unable to lift obstacles related to the opacity of 'black box' and other complex algorithms, impinging on their ability to challenge AI systems that infringe on their own fundamental rights.

For example, in European discrimination litigation, the burden of proof can shift to the defendant if applicants are able to establish a *prima facie* case of discrimination, that is, 'facts from which it may be presumed that there has been ... discrimination'.⁵³ Yet, it has been argued that bringing such preliminary evidence can amount to a major hurdle for individual applicants.⁵⁴ Elevated to human rights, IPRs could be used by defendants as a basis for refusing to provide information on the parameters of algorithmic decision-making. This could, in turn, impede not only end users', but also auditors', and even judges' efforts to identify potential discriminatory features in an AI system.⁵⁵ A human right to property applied to AI

⁵² See Robin Gross, 'Information Property Rights and the Information Commons' in Rikke Frank Jørgensen, Ernest J Wilson, and William J Drake (eds), *Human Rights in the Global Information Society* (MIT Press 2006) 110. See also the chapter by Shu Li, Béatrice Schütte, and Lotta Majewski in this volume.

⁵³ See Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin [2000] OJ L180/22, art 8; Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation [2000] OJ L303/16, art 10; Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services [2004] OJ L373/37, art 9; Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) [2006] OJ L204/23, art 19. See also *Nachova and Others v Bulgaria* [GC] App nos 43577/98 and 43579/98 (ECtHR, 6 July 2005), para 147.

⁵⁴ See eg Janneke Gerards and Raphaële Xenidis, *Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Anti-Discrimination Law* (European Commission 2020) 73–75, 144–45.

⁵⁵ In *Danfoss*, the ECJ established that the complete lack of transparency of a given system can justify a shift of the burden of proof once *prima facie* evidence has been adduced. This principle can effectively mitigate end users' difficulty accessing information to establish a *prima facie* case of discrimination. See

systems might therefore considerably increase the evidentiary burden of potential victims. In that case, detecting algorithmic discrimination would only remain possible through systematically testing algorithmic output *a posteriori*, a costly solution that individual applicants will not be able to avail.

Another example of such clashes is the right to privacy, private life, and the protection of personal data, which could be endangered by the application of a human right to property in the context of AI systems. For example, it has been argued that some platforms share users' personal data with commercial partners without offering transparency on how, why, and with whom such data is shared.⁵⁶ If claimed, the right to IP could support such platforms' claims against disclosing to end users information about how, to whom, and what kind of personal data is shared by their AI systems. This would heighten the threshold for applicants willing to bring a legal claim to enforce their right to privacy.

In the context of the right to a fair trial and due judicial process and the fundamental right to an effective remedy, the infamous example of the COMPAS risk assessment system used in the US has shown that the lack of transparency and the failure to properly explain how algorithmic decision support systems work is liable to lead to severe breaches.⁵⁷ The enforcement of the fundamental right to liberty and security might also be jeopardised if IPRs compromise citizens' ability to gain information on AI systems used for face recognition, crime detection, and public surveillance overall. A similar logic applies in the context of other rights such as the right to education, where algorithmic decision support or decision-making systems are used to assess candidates' applications or to allocate grades, and where IPRs might prevent concerned subjects from gaining insights into the systems' functioning and thereby from raising complaints.⁵⁸ Proprietary rights might also raise enforcement issues in the context of the right to health, if inscrutable AI systems are used to support harmful decisions about patients' health.⁵⁹ In sum, the right to property applied to AI could sharpen black box problems and the serious risks they pose in relation to the exercise of numerous fundamental rights.

At the same time, the right to property is not absolute and in case of clashes with other human rights, several limitations could apply. The first kind of limits

Case C-109/88 *Handels- og Kontorfunktionærernes Forbund I Danmark v Dansk Arbejdsgiverforening*, acting on behalf of *Darfoss* EU:C:1989:383, para 16. However, it has been argued that this principle does not readily apply in the realm of recruitment procedures, see Ljupcho Grozdanovski, 'In Search of Effectiveness and Fairness in Proving Algorithmic Discrimination in EU Law' (2021) 58(1) *Common Market Law Review* 99 115–18.

⁵⁶ Frank Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information* (Harvard UP 2015) 143–45.

⁵⁷ See *Loomis v Wisconsin* 881 N.W.2d 749 (Wis 2016), cert. denied, 137 S.Ct. 2290 (2017).

⁵⁸ See eg Ryan S Baker and Aaron Hawn, 'Algorithmic Bias in Education' (2022) 32 *International Journal of Artificial Intelligence in Education* 1052.

⁵⁹ See eg Ziad Obermeyer and others, 'Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations' (2019) 366(6464) *Science* 447.

purports to the so-called ‘right to explanation’ that commentators had initially hoped to draw from the GDPR to help data subjects request information about the rationale of specific algorithmic decisions. However, Wachter and others have argued that the GDPR only guarantees a much more limited ‘right to be informed’ about ‘the usage and functionality of automated decision-making methods’ in general.⁶⁰ Hence, even though the Data Protection Working Party has clarified in relation to the right of access (article 15 of the GDPR) that ‘controllers cannot rely on the protection of their trade secrets as an excuse to deny access or refuse to provide information to the data subject’,⁶¹ an elevated human right status for IPRs under the purview of the right to property risks hampering data subjects’ right to access ‘meaningful information about the logic involved’ in automated decisions.⁶² Such an ‘intensified protection of intellectual property’ might encourage value-related claims by businesses that might, in turn, ‘severely crippl[e] the right to explanation’ and therefore leave ‘citizens subjected to momentous scoring or profiling algorithms … in the dark as to the very reasons for outcomes affecting their destinies’.⁶³

The second type of limitations that apply in case of clashes between human rights is the balancing exercise that courts will have to conduct to test the proportionality of existing infringements. In this case, transparency requirements might take several forms depending on the outcome of the proportionality test. Pasquale, for example, suggests that such transparency requirements can be modelled according to three questions: ‘How much does the black box firm have to reveal? To whom must it reveal it? And how fast must the revelation occur?’⁶⁴ Tutt, in turn, has identified a transparency scale ranging from—on its more restrictive end—exclusive disclosure of limited elements of an AI system (eg code and training data) to third-party certifiers to—on its more extensive end—a disclosure of technical specifications and a pre-emption of trade secrets.⁶⁵ Balancing out the right to property with other human rights will certainly require ‘effective transparency’⁶⁶ or what Pasquale calls ‘qualified transparency’, that is ‘limiting revelations in order to respect all the interests involved in a given piece of information’.⁶⁷ In addition, it will be crucial for potential victims of human rights breaches that the information

⁶⁰ Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’ (2017) 7(2) International Data Privacy Law 76, 90.

⁶¹ Article 29 Data Protection Working Party, ‘Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679’ (2018) <<https://ec.europa.eu/newsroom/article29/items/612053/en>>.

⁶² *ibid.*

⁶³ Paul B de Laat, ‘Algorithmic Decision-making Employing Profiling: Will Trade Secrecy Protection Render the Right to explanation Toothless?’ (2022) 24 *Ethics and Information Technology* 1, 17.

⁶⁴ Pasquale (n 56) 142.

⁶⁵ See Andrew Tutt, ‘An FDA for Algorithms’ (2017) 69(1) *Administrative Law Review* 83, 110.

⁶⁶ David Leslie and others, *Artificial Intelligence, Human Rights, Democracy, and The Rule of Law: A Primer* (Council of Europe and The Alan Turing Institute 2021) 38.

⁶⁷ Pasquale (n 56) 142.

provided to them is intelligible. Hence, in certain cases such as AI relying on complex ML systems, transparency requirements will not be sufficient to avoid abuses of IPRs and explainability will be key.⁶⁸

This section has reviewed how the right to property, applied to AI systems, could affect—both positively and negatively—the exercise of other human rights. Yet, further conundrums could arise if future legal reforms open the door to recognising legal personhood to AI. While this question lies beyond the scope of this chapter,⁶⁹ it is important to recognise that granting IP protection to an AI-generated output might then require treating AIs as inventors.⁷⁰ If IPRs are elevated to fundamental rights, that scenario would grant AIs the status of human rights subjects entitled to a protection of their property. This would in turn trigger new and complex fundamental rights clashes and questions.

4 Contemporary Regulatory Solutions Aimed at Mitigating the Negative Effects and Facilitating the Positive Ones

International IP treaties devote little attention to human rights considerations.⁷¹ This is also evident while looking at standards for the implementation of exceptions to IP rules, which need to be ‘necessary’ for protecting the public interest. The approach shall be the least restrictive, as the term is interpreted as mandating contracting parties to rely on the measure which might cause least harm to the interests of right holders.⁷² Yet, exceptions can play an important role in easing the tensions between countervailing rights.

4.1 IP Exceptions: A Tool to Solve the Tension?

The IP law framework provides for exceptions and limitations to exclusive rights in order to strike an appropriate balance between the interests of right holders, third parties, and the public. Although they are not regarded as proprietary rights *stricto sensu*, the analysis focuses on trade secrets for they are the most widely used tool to protect AI systems. This sub-section sheds light on whether, in addition to human

⁶⁸ Leslie and others (n 66) 38.

⁶⁹ On this issue, see also the chapter by Klaus Heine in this volume.

⁷⁰ In this regard see, among others Maria Iglesias, Sheron Schamiulia, and Amanda Anderberg, ‘Artificial Intelligence and Intellectual Property: A Literature Review’ (Publication Office of the European Union 2021) 1, 12–19 <<https://publications.jrc.ec.europa.eu/repository/handle/JRC119102>>.

⁷¹ See section 2.1.

⁷² H Grosse Ruse-Khan, ‘Assessing the Need for a General Public Interest Exception in the TRIPS Agreement’ in A Kur (ed), *Intellectual Property in a Fair World Trade System* (Edward Elgar 2011) 167, 173.

rights protection as such, the use of exceptions and limitations can represent a viable tool to mitigate the clashes between right holders' interests and transparency and access rights. In order to gain more insights into the implementation of international standards it is relevant to analyse the legislation at regional and national level. To this aim, the EUTSD and related interpretation is taken as an example.

Article 1(2) of the EUTSD clarifies that the implementation of the EUTSD shall not affect the application of EU or national rules which require trade secrets' holders 'to disclose, for reasons of public interests, information, including trade secrets, to the public or to administrative and judicial authorities for the performance of the duties of those authorities'.⁷³ Although transparency and access rights are therefore safeguarded in these circumstances, the provision clearly does not cover several other instances in which disclosure may be necessary, including all disclosures between private entities. The legislator has therefore provided for a list of limitations and exceptions to trade secrets protection.⁷⁴ In particular, article 3 of the EUTDS lists the circumstances in which the acquisition of trade secrets is considered lawful. These include: (a) independent discovery or creation; (b) reverse engineering; (c) exercise of the right of workers or workers' representatives to information and consultation; and (d) any other practice which is in conformity with honest commercial practices. In addition, article 5, titled 'Exceptions', mandates that member states 'shall ensure' exceptions to trade secrets protection in the cases in which the alleged acquisition, use or disclosure of the trade secret occurred for: (a) the exercise of the right to freedom of expression and information; (b) for revealing misconduct, wrongdoing, or illegal activity, provided that the respondent acted for the purpose of protecting the public interest; (c) disclosure by workers to their representatives for the legitimate exercise of their functions; and (d) the protection of a legitimate interest recognised by EU or national law.⁷⁵ A detailed analysis of the benefits and flaws of each of the listed limitations and exceptions goes beyond the scope of the present contribution.⁷⁶ Suffice here to say that although they represent a welcome attempt of the legislator to ensure the right balance between secrecy, exclusive rights, transparency, and access rights in a broad range of situations, their implementation presents challenges. For example, and of relevance in this context, the 'public interest' exception can be raised only if the acquisition, use, or disclosure was 'for revealing misconduct, wrongdoing or illegal activity'.⁷⁷ In other words, the public interest is not regarded as a stand alone defence,⁷⁸ and as a consequence, access and transparency are not exempted in several

⁷³ EUTSD, art 1(2)(b).

⁷⁴ EUTSD, arts 3, 5.

⁷⁵ EUTSD, art 5.

⁷⁶ For a detailed analysis in this regard, see Tania Aplin, 'The Limits of EU Trade Secret Protection' in S Sandeen, C Rademacher, and A Ohly (eds), *Research Handbook on Information Law and Governance* (Edward Elgar 2021) 1, 18.

⁷⁷ EUTSD, art 5(b).

⁷⁸ Noto La Diega (n 44) 13.

deserving circumstances. Of course, the scope of application of such an exception depends on how the wording has been transposed and will be interpreted at national level. Yet, many member states appear to follow the approach of the EU legislator, not providing for public interest as a defence in its own right.⁷⁹ As a result, exceptions and limitations to IPRs represent a vital instrument for overcoming the clashes between different fundamental rights and interests, but certain flaws and shortcomings can undermine their deployment and effectiveness.

4.2 Policy Proposals that Could Improve the Current Status Quo

In light of existing shortcomings, it is recommended to favour a broader interpretation of IP exceptions. For example, as mentioned in relation to trade secrets, it is advisable to amend, interpret and address the ‘public interest’ exception as a defence in its own right.

In addition, in the context of future improvements to the current legal framework, it is important to take into consideration the relevance of exceptions to IPRs to balance the different interests reflected in the wording of the new draft EU Artificial Intelligence Act (AI Act).⁸⁰ While providing for a framework for trustworthy artificial intelligence (AI), the legislator clarifies that ‘[n]ational competent authorities and notified bodies involved in the application of this Regulation shall respect the confidentiality of information and data obtained in carrying out their tasks and activities’ in order to protect IPRs, and confidential information or trade secrets, except when the instances referred to in article 5 of the EUTSD apply.⁸¹ The attempt to reconcile the right of access and transparency without hampering IPRs in a disproportionate manner is noteworthy.⁸² However, the coordination between the AI Act and article 5 of the EUTSD, as currently proposed, may well not sufficiently serve as a safeguard for the exercise of the right to access and transparency in the public interest, at the dawn of the widespread deployment of AI systems. First of all, the referred provision presupposes a ‘misconduct, wrongdoing and illegal activity’. Second, the AI Act limits the right to access relevant documentation and information only to public authorities, thereby overlooking the role that individuals and private entities such as non-governmental organisations (NGOs)

⁷⁹ Letizia Tomada, ‘Intellectual Property Rights at the Intersection of Algorithmic Bias Examination and Correction in the proposed EU Artificial Intelligence Act’ (on file with the author).

⁸⁰ In April 2021, the European Commission presented a proposal for a ‘Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts’ (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>. The proposal aims at avoiding ‘significant risks to the health and safety or fundamental rights of persons’: see Explanatory Memorandum, p 4; and AI Act, Recitals 5, 13, 14.

⁸¹ AI Act, art 70(1)(a).

⁸² See also Explanatory Memorandum (n 80) 11.

can have in unveiling public interests' concerns, such as bias examination and correction.⁸³

All in all, the existence and implementation of IPRs' exceptions can tip the balance between countervailing fundamental rights and interests. However, much room is left for future improvement, both as regards flaws and discrepancies in their scope of application among different member states and as regards the coordination with other relevant legal instruments.

5 Conclusion

This chapter has shown how recognising the human rights dimension of IP in the context of AI applications has several implications. On the one hand, it could foster investments and innovation in the field, with positive ramifications on economic freedoms. On the other, however, it is liable to create severe clashes with other fundamental rights. By restricting opportunities for access, information, and transparency, the recognition of a human rights dimension to IP in the AI context creates direct conflicts with related rights such as the freedom of expression and information. Moreover, this can indirectly affect the exercise of other human rights where the lack of transparency and the lack of access to meaningful information about AI systems seriously hampers their effective enforcement. Across diverse contexts, the legislator has acknowledged the existence of this tension and has provided for regulatory solutions aimed at mitigating related risks, such as exceptions to IP protection within the IP legal framework. Yet, their application and coordination with other relevant existing and upcoming legal instruments presents limitations and challenges. A broader and more uniform interpretation of specific IP law exceptions and tailored regulatory adjustments can, therefore, be key to achieving an appropriate balance between relevant fundamental rights.

⁸³ Tomada (n 79); see also G Spina Alí and R Yu, 'Artificial Intelligence between Transparency and Secrecy: From the EC Whitepaper to the AIA and Beyond' (2021) 12 European Journal of Law and Technology 1, 17; N Mehrabi and others, 'A Survey on Bias and Fairness in Machine Learning' (2021) 54(6) ACM Computing Surveys 1–35.

PART III

ARTIFICIAL INTELLIGENCE
AND PRIVACY

Artificial Intelligence and the Right to Privacy

*Alessia Zornetta and Ignacio Cofone**

1 Introduction: The Rights to Privacy and Data Protection

The concept of privacy can be interpreted in several ways. In Daniel Solove's famous taxonomy, there are six privacy conceptualisations.¹ First, privacy can be seen as the right to be let alone, free from external constraints such as state interference.² Second, privacy is recognised as limited access to the self, affording individuals the right to reduce the access to and usage of information about them.³ Third, privacy can be interpreted as secrecy, meaning that individuals' privacy is violated when matters they wish to keep confidential are disclosed.⁴ Fourth, privacy can be viewed as people's control over their personal information, related to self-determination since it allows individuals to decide what information about them is shared with others.⁵ Fifth, privacy can be understood as the right to protect one's personhood by enabling people to develop their identity and make choices free from constraints.⁶ Finally, privacy can be interpreted as intimacy, stressing its importance for human relationships.⁷

Since the nineteenth century, the right to privacy has been recognised in national constitutions, supranational organisations, and instruments of international law. The right to privacy is recognised globally. It is present, for example, in article

* We thank Alberto Quintavalla and Jeroen Temperman for their helpful comments on the chapter.

¹ Daniel J Solove, 'Conceptualizing Privacy' (2002) 90 California Law Review 1087, 1102–124. See also Woodrow Hartzog, 'What is Privacy? That is the Wrong Question' (2021) 88 University of Chicago Law Review 1677.

² Samuel D Warren and Louis D Brandeis, 'The Right to Privacy' (1890) 4 Harvard Law Review 193.

³ Herman T Tavani, 'Philosophical Theories of Privacy: Implications for an Adequate Online Privacy Policy' (2007) 38 *Metaphilosophy* 1.

⁴ Benjamin E Hermalin and Michael L Katz, 'Privacy, Property Rights and Efficiency: The Economics of Privacy as Secrecy' (2006) 4 Quantitative Marketing and Economics 209; Christine S Scott-Hayward, Henry F Fradella, and Ryan G Fischer, 'Does Privacy Require Secrecy: Societal Expectations of Privacy in the Digital Age' (2015) 43 American Journal of Criminal Law 19.

⁵ Alan Westin, *Privacy and Freedom* (Ig 1967) 4–6.

⁶ Philip E Agre and Marc Rotenberg, 'Technology and Privacy: The New Landscape' (1997) 11 Harvard Journal of Law & Technology 3.

⁷ Avrum G Weiss, 'Privacy and Intimacy: Apart and a Part' (1987) 27 J Humanist Psychol 1; Charles Fried, 'Privacy' (1968) 77 Yale Law Journal 475.

12 of the Universal Declaration of Human Rights (UDHR) and article 17 of the International Covenant on Civil and Political Rights (ICCPR) and, within Europe, in article 7 of the European Union (EU) Charter of Fundamental Rights and article 8 of the European Convention on Human Rights (ECHR). It poses both negative and positive obligations on the state: the state must refrain from infringing individuals' privacy and ensure that individuals' privacy is protected against violations.

Over the years, the European Court of Human Rights (ECtHR) has interpreted the provisions of article 8 of the ECHR to protect privacy⁸ and has assessed state action's compliance with article 8's legitimacy, legality, and proportionality requirements.⁹ In jurisdictions where the right to privacy is not explicitly stated, courts have interpreted constitutional provisions and case law to afford it constitutional protection. For instance, in the United States (US), the Supreme Court has interpreted the Fourth Amendment of the US Constitution to include the right to privacy.¹⁰

With the rise of the first databases and data processors in the 1970s, scholars started to point to the risks linked to the increasingly frequent use of technologies that infringe on people's privacy rights. As a response, the concept of data protection was developed and gained traction among regulators and judiciaries as an extension of the right to privacy.¹¹

Data protection refers to the protection of information of an identifiable natural person. Examples of identifiable information are names, dates of birth, photographs, videos, e-mail addresses, and IP addresses. Because of their close relationship, most European Union jurisdictions have interpreted the right to privacy to give origin to the right to data protection. The ECtHR has ruled that article 8 also awards constitutional safeguards to personal data.¹² The EU is unique in that the right to data protection and the right to privacy are protected by separate provisions. Data protection is safeguarded in article 8 of the EU Charter and article 16 of the Treaty on the Functioning of the European Union (TFEU). This has allowed the EU to introduce rules specific to data protection, first in the Data Protection Directive and then in the General Data Protection Regulation (GDPR),¹³ which

⁸ ECHR, 'Guide on Article 8 of the European Convention on Human Rights' (2021), 48–66.

⁹ Eg *Roman Zakharov v Russia* App no 47143/06 (ECtHR, 4 December 2015).

¹⁰ *Katz v United States* 389 US 347 (1967); *Griswold v Connecticut* 381 US 479 (1965); and *Kyllo v United States* 533 US 27 (2001).

¹¹ Case C-291/12 *Michael Schwarz v Stadt Bochum* [2014] ECR 670; Joined Cases C-293/12 and C-594/12 *Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others* [2014] ECR 238.

¹² *Amann v Switzerland* App no 27798/95 (ECtHR, 16 February 2000), para 65; *Rotaru v Romania* App no 28341/95 (ECtHR, 4 May 2000), para 43.

¹³ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (GDPR).

have served as starting point for several jurisdictions worldwide.¹⁴ The GDPR has been one of the most influential instruments of EU law.¹⁵

Although both rights are necessarily interwoven, they are not identical. The right to privacy includes interpersonal relations offline that fall outside of the scope of data protection. The right to data protection, on the other hand, not only protects privacy: it also provides protections for situations in which personal data are collected, processed, or shared, and individuals' privacy interests are not infringed.

The remainder of the chapter presents privacy and data protection principles applicable to AI and then delves into each of the various principles, analysing them and presenting issues that arise when they are applied to AI as currently developed, and how complying with each principle would ensure the respect for the human right to privacy. We focus on EU law as it is the most influential and comprehensive body of law in this area but our analysis is also applicable to other jurisdictions that incorporate these principles, such as those jurisdictions seeking or having adequacy status with regards to the GDPR.¹⁶

2 Applicability of Privacy and Data Protection Principles

The GDPR principles and other data protection principles were highly influenced by pre-existing guidelines.¹⁷ Although legislations differ, almost all reflect the OECD's 'Guidelines governing the protection of privacy and trans-border flows of personal data', also known as the 'Fair Information Practice Principles' (FIPPs), which aim to ensure fair processing of personal data.¹⁸ Because artificial intelligence (AI) often deals with personal data, principles of privacy and data protection closely impact the development and use of AI.

Not all uses of AI will trigger the application of the GDPR—and its equivalent data protection legislations outside of the EU. Developers only need to fulfil the duties provided by the GDPR if they engage in processing personal data,¹⁹ that

¹⁴ Eg Andorra's Personal Data Protection Act 2021; Argentina's Personal Data Protection Act 2000; Canada's Personal Information Protection and Electronic Documents Act 2000; Israel's Protection of Privacy Law (as amended 2007); Japan's Act on the Protection of Personal Information (as amended 2020); and New Zealand's Privacy Act 2020.

¹⁵ Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius, 'The European Union General Data Protection Regulation: What It Is and What It Means' (2019) 28 *Information & Communications Technology Law* 65; Guy Aridor, Yeon-Koo Che, and Tobias Salz, 'The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR' (2020) National Bureau of Economic Research 26900/2020.

¹⁶ See EU Commission, 'Adequacy Decisions' (2021) <https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en>. Once a jurisdiction is granted adequacy status, data can be transferred between the EU and such jurisdiction without further safeguards.

¹⁷ Hoofnagle, Van der Sloot, and Frederik Zuiderveen (n 15) 70.

¹⁸ OECD, *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data* (OECD Publishing 2002) 11–18.

¹⁹ GDPR, art 2(1).

is, processing ‘any information relating to an identified or identifiable natural person’.²⁰

Therefore, developers must ask three questions: (i) in the collection stage, are we collecting personal data? (ii) are we using personal data to train, test, or validate our model? and (iii) are we inputting or inferring personal data with our model? If the answer to any of these questions is yes, meaning personal data is used at any stage of the process, then privacy and data protection principles apply.

There are two key actors with different levels of responsibility in case of infringement. The first key actor is the controller, who must comply with all GDPR principles and rules.²¹ The controller is the individual or entity who determines the purpose and the means of the processing. For instance, in a facial recognition app, the controller will decide why the processing is happening and the means to be used in the processing. That is, which software will be used, which data will be processed, whose data will be collected, and who will benefit from the processing. In AI processing, different organisations might be involved at different processing stages and for different purposes. In that case, they would all be controllers: when two entities’ decisions are necessary for processing to take place and they both pursue the same purpose or two closely linked ones, the GDPR contemplates the possibility of the entities being simultaneously recognised as controllers and share the consequent responsibilities—a phenomenon called joint controllership.²²

The second key actor is the data processor. A processor is the person or entity who is entrusted with processing personal data on behalf of the controller, following their instructions. This means that the processor cannot undertake processing for a purpose other than that established by the controller. For instance, in the facial recognition example, the developer of the software used will be a processor. The processor must comply with a subset of the controller’s obligations.²³

The main principles related to the processing of personal data under the GDPR—and many of the data protection laws based on the GDPR—that controllers and processors must keep in mind when developing AI are: lawfulness, fairness, and transparency; purpose limitation; data minimisation and storage limitation; accuracy; integrity and confidentiality; and accountability.²⁴ These principles are also recognised in international legal frameworks including the OECD Data Protection Principles, the Council of Europe’s Convention 108, and in jurisdictions such as Canada, Brazil, Peru, and Japan.²⁵ The principles can clash with some uses of AI since, to obtain robust results, as much high-quality data as possible needs to be

²⁰ GDPR, art 4(1). See also the chapter by Andrea Pin in this volume.

²¹ GDPR, art 4(7).

²² See European Data Protection Board, *Guidelines 07/2020 on the Concepts of Controller and Processor in the GDPR* (2020) 18–20; GDPR, art 26(1).

²³ GDPR, art 4(8). Eg security, maintain a record of processing, etc.

²⁴ GDPR, art 5.

²⁵ See n 14; Council of Europe, Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, CETS No. 108, (1981) (Convention 108).

collected, processed, and stored. This clash has even led some scholars to question the feasibility of maintaining these principles applicable to AI.²⁶

3 Lawfulness and Fairness in AI

A requirement for processing personal data under the FIPPs is that it be ‘lawful, fair, and transparent’.²⁷ In the GDPR, processing is deemed lawful if undertaken based on one of the following grounds: ‘consent’, ‘performance of a contract’, ‘compliance with a legal obligation’, ‘protection of the vital interests of the data subject’, ‘public interest’, and ‘legitimate interests’.²⁸ The principle of lawfulness also requires that controllers’ data processing complies with the rule of law, meaning that they must consider the individual’s legitimate expectations and be subject to independent oversight.²⁹

Fulfilling the lawfulness requirements in AI applications might not be straightforward. Certain grounds are available only for a specific phase of the AI processing rather than its entirety. Development and deployment may have to rely on different grounds. For instance, relying on ‘performance of a contract’ might be appropriate when using AI processing to provide an estimate of the costs of a service. But it might be inappropriate as a ground to develop such an AI model in the first place, as the British enforcement authority seems to indicate.³⁰ Similarly, the ‘vital interest of the data subject’ justification is likely to be lawful only in case of emergency medical diagnosis of unconscious patients, meaning it works only in the deployment phase.

Further limitations apply when AI models are used in automated individual decision-making and profiling practices.³¹ According to article 22 of the GDPR—and analogous articles in other privacy and data protection legislations—only three grounds allow automated decision-making to be lawful, namely if: (i) necessary for the performance of a contract; (ii) authorised by a law to which the controller is subject; or (iii) based on the individual’s explicit consent.³² This is the case also, for example, in Brazil and Uruguay.³³ In these cases, to comply with the

²⁶ Eline Chivot and Daniel Castro, ‘The EU Needs to Reform the GDPR to Remain Competitive in the Algorithmic Economy’ (*Center for Data Innovation*, 13 May 2019); and Tal Zarsky, ‘Incompatible: The GDPR in the Age of Big Data’ (2017) 47 Seton Hall Law Review 1009.

²⁷ OECD, ‘Collection Limitation Principle’ (n 18).

²⁸ GDPR, art 6.

²⁹ Case C-465/00 *Rechnungshof v Österreichischer Rundfunk* [2003] ECR I-04989.

³⁰ UK Information Commissioner’s Office (ICO), ‘Guidance on AI and Data Protection’ (ICO, 30 July 2020) <<https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/>>.

³¹ See the chapter by Helga Molbæk-Stensig and Alexandre Quemy in this volume.

³² GDPR, art 22(2).

³³ See eg Lei no 13.709 de 14 de agosto de 2018: Dispõe sobre a proteção de dados pessoais (LGPD), 20 (BRA); Ley no 18.331 de Protección de Datos Personales y Acción de Habeas Data (LPDP) de 11 de agosto de 2008, 16 (URY).

lawfulness principle, controllers must ensure that individuals are informed about the processing, that they can request human intervention and challenge the decision, and that the system's accuracy is assessed regularly.³⁴

The fairness requirement implies that personal data should not be processed in a way that goes against data subjects' reasonable expectations as to how the data will be used or processed in a way that unjustifiably adversely impacts them.³⁵ Unjustifiably means that the controller and the processor must conduct balancing and proportionality tests, assessing whether the risks are sufficiently mitigated and worth undertaking to achieve the purpose of the collection and processing. In machine learning (ML), the principle of fairness is particularly relevant in two ways: accuracy and non-discrimination.

Complying with the fairness requirement implies that controllers ensure that AI models are statistically accurate.³⁶ That is, that the outputs of the model remain coherent and correct. Although it is impossible to ensure complete accuracy, to comply with the GDPR controllers must take necessary steps to reasonably minimise inaccuracies and assess the impact that such inaccuracies might have on data subjects.³⁷

Fairness also requires that AI predictions do not result in discrimination against data subjects.³⁸ Algorithmic bias could compromise the model's compliance with the principle of fairness if it leads to discriminatory outcomes. The consistency with which ML systems make decisions can lead to unjustifiable adverse results when the datasets used in the training phase contain bias, when people involved in the process translate their bias into the system, or when the system uses a biased output variable.³⁹ When bias is present in training datasets, there is a risk that the AI model unfairly discriminates against individuals. For instance, considering the facial recognition example mentioned above, bias has been shown to exist towards non-white subjects because they were underrepresented in the training phase.⁴⁰

³⁴ ibid; GDPR, art 22(2).

³⁵ Gianclaudio Malgieri, 'The Concept of Fairness in the GDPR' (Conference on Fairness, Accountability, and Transparency, 2020) 156.

³⁶ GDPR, Recital 71; Philipp Hacker, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law' (2018) 55 Common Market Law Review 1143–85.

³⁷ Giovanni Sartor and Francesca Lagioia, 'The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence' (EU Parliament PE641.530, 2020) 44–45. See also ICO (n 30).

³⁸ See also Part IV of the present volume.

³⁹ Ignacio Cofone, 'Algorithmic Discrimination is an Information Problem' (2019) 70 Hastings Law Journal 1389.

⁴⁰ Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) 81 Proceedings of Machine Learning Research 1–15 <<https://perma.cc/ZPM9-TA23>>. For additional concerns regarding facial recognition technologies (FRT), see the chapters by Natalia Menéndez González; and Malcolm Katrak and Ishita Chakrabarty in this volume.

4 Transparency in AI

Personal data processing must be transparent, meaning that data subjects have the right to be informed of the processing in an easily accessible and understandable manner.⁴¹

The transparency principle, stipulated in the GDPR and other jurisdictions such as Canada, centrally affords data subjects the right to information and access.⁴² Accordingly, data subjects have the right to obtain specific information from the data controller, including a confirmation of ongoing processing, the personal data being used, and details of the processing, including its risks and adopted safeguards. The aim of transparency is to enhance processing systems' accountability and improve data subjects' understanding of the process. Thus, transparency also implies that data subjects can obtain an explanation of how and why the processing led to a decision about them.⁴³

Regarding AI, transparency rights empower data subjects to obtain information about the models processing their data. This means that individuals could use them to understand how, for instance, their credit score was calculated. However, for most AI models, providing this kind of information can be complicated—and, for some, even undesirable.⁴⁴ Some applications are efficient because they can process data almost independently of their programmers. This means that even the programmers might not be able to explain a decision. This issue has been described as the *black box phenomenon* and concerns most deep artificial neural networks applications.⁴⁵ A neural network consists not only of an input and an output layer but also multiple hidden layers of artificial neurons. These have been labelled the black box phenomenon because it is difficult to understand how the result has been reached when data is processed through the network.

Transparency is relevant when assessing AI applications' accountability. However, the opacity of AI models has led some to argue that legal accountability mechanisms are not sufficiently developed to keep pace with AI.⁴⁶ The processing

⁴¹ EU Commission, 'Article 29 Working Party Guidelines on transparency under Regulation 2016/679' (2018) <<https://perma.cc/N87M-ERKG>>.

⁴² GDPR, arts 13–15; LGPD, arts 6(VI), 9; Personal Information Protection and Electronic Documents Act, SC 2000, c 5, ss 4.8–4.9 (Canada) (PIPEDA).

⁴³ Margot Kaminski, 'The Right to Explanation, Explained' (2019) 4 Berkeley Technology Law Journal 1.

⁴⁴ See eg Paul Ohm, 'Changing the Rules: General Principles for Data Use and Analysis' in Julia Lane and others (eds), *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (CUP 2014) 100; Ignacio Cofone and Katherine Strandburg, 'Strategic Games and Algorithmic Secrecy' (2019) 64(4) McGill Law Journal 623.

⁴⁵ See International Working Group on Data Protection in Telecommunications, 'Working Paper on Privacy and Artificial Intelligence' (2018) (AIWP).

⁴⁶ Mike Ananny and Kate Crawford, 'Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973–89; Reuben Binns, 'Algorithmic Accountability and Public Reason' (2018) 31 Philosophy & Technology 543–56. See also the chapter by Klaus Heine in this volume.

of data by some AI applications cannot be fully explainable. This impacts the ability to independently audit the functioning of such applications to ensure their compliance with data protection principles. The unintelligibility and opacity of some AI systems, which interferes with transparency, might also interfere with fairness and accountability requirements if neither individuals nor supervisory authorities have means of knowing whether the decisions made through an opaque process were accurate and fair.⁴⁷

In AI, compliance with the right to be informed might be incompatible or hindered by compliance with other data protection principles. For instance, data minimisation might require training data to be modified and removed of any contact information. In this case, it might be difficult, if not impossible, to communicate information to the data subjects involved.⁴⁸

5 Purpose Limitation in AI

Another relevant principle of data protection is purpose limitation.⁴⁹ According to it, data can only be processed for the purposes explicitly specified to data subjects at the moment of collection (independent of the legal basis used) or for further purposes deemed compatible with the original purpose.⁵⁰

In AI applications, compliance with purpose limitation might pose challenges to controllers and processors.⁵¹ First, the development and deployment phases have significantly different purposes (one for design and training, the other for actual use). For instance, in the facial recognition example, individuals might consent to have their data used to train an AI model to recognise faces. But this does not mean that data subjects also consented to their data being used in systems aimed at crime prevention, authentication software, or friends tagging on social media.⁵²

⁴⁷ See generally Tal Zarsky, ‘The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making’ (2016) 41 *Science, Technology, and Human Values* 118.

⁴⁸ ICO (n 30).

⁴⁹ GDPR, art 5(1)(b); OECD, ‘Purpose Specification’ and ‘Use Limitation’ (n 18); Council of Europe, Recommendation R (87) 15 on Use of Personal Data in the Police Sector, 2.1 and 4; Convention 108, art 5(b)–(c); GDPR, art 6(4).

⁵⁰ See Merel Elize Koning, ‘The Purpose And Limitations Of Purpose Limitation’ (DPhil Thesis, Radboud University Nijmegen 2020); EU Commission, ‘Can We Use Data for Another Purpose?’ (European Commission, 2021) <<https://perma.cc/UUL7-47DP>>. Cf CCPA, para 1798.100(b); PIPEDA, s 6 Sch 1 4.5; LGPD, art 6(III); Act on the Protection of Personal Information (No 57 of 2003, as amended 2020) (Japan) (APPI), art 16.

⁵¹ Ignacio Cofone, ‘Policy Proposals for PIPEDA Reform to Address Artificial Intelligence’ (Office of the Privacy Commissioner of Canada, November 2020) <https://www.priv.gc.ca/en/about-the-opc/what-we-do/consultations/completed-consultations/consultation-ai/pol-ai_202011/>.

⁵² See eg ICO (n 30); see also UK Information Commissioner’s Office, ‘Guide to the General Data Protection Regulation’ (1 January 2021) 25–28 <<https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr-1-1.pdf>>.

The purpose limitation principle can be particularly challenging when applied to AI processing, as legal scholars and developers pointed out.⁵³ By requiring that all legitimate purposes for data processing are specified at the moment of collection by the controller, the principle imposes substantial limitations on most AI uses. This is because a characteristic of ML models is that developers themselves cannot always predict what will be the purpose of the processing.⁵⁴

Because controllers are prohibited from drafting too broad definitions of purposes, the principle might require controllers to constantly re-obtain data subject's consent (or a different legitimising basis) to ensure that further processing remains compliant.⁵⁵ Yet, the GDPR and similar data protection laws (eg those of India, Peru, and Canada) provide an exception when further processing regards scientific or statistical research.⁵⁶ In such cases, data controllers and processors can further process personal data without re-obtaining data subjects' consent because further processing is considered compatible. AI is included in the GDPR statistical research exception only when it does not involve profiling.⁵⁷ Despite these considerations, AI's unforeseeability and unpredictability remain a challenge.⁵⁸

6 Data Minimisation and Storage Limitation in AI

The data minimisation and storage limitation principles are related in how they apply to AI. The data minimisation principle requires that personal data is adequate and relevant to, and only collected and processed to the extent necessary for, the purposes specified.⁵⁹ This requires that controllers and processors consider all

⁵³ Eg Zarsky (n 26) 1006; Ignacio Cofone, 'Beyond Data Ownership' (2021) 43 Cardozo Law Review 501, 559–64.

⁵⁴ Nikolaus Forgó, Stefanie Hänold, and Benjamin Schütze, 'The Principle of Purpose Limitation and Big Data' in Marcelo Corrales, Mark Fenwick, and Nikolaus Forgó (eds), *New Technology, Big Data and the Law* (Springer 2017) 34; Luca Marelli, Elisa Lievevrouw, and Ine Van Hoyweghen, 'Fit for Purpose? The GDPR and the Governance of European Digital Health' (2020) 41 Policy Studies 453–54; Robin Pierce, 'Machine Learning for Diagnosis and Treatment' (2018) 4 European Data Protection Law Review 340; cf Michèle Finck and Asia Biega, 'Reviving Purpose Limitation and Data Minimisation in Personalisation, Profiling and Decision-Making Systems' (2021) Technology and Regulation 44–61 <<https://perma.cc/KS4L-BVWB>>.

⁵⁵ Liana Colonna, 'Data Mining and its Paradoxical Relationship to the Purpose of Limitation' in Serge Gutwirth, Ronald Leenes, and Paul de Hert (eds), *Reloading Data Protection: Multidisciplinary Insights and Contemporary Challenges* (Springer 2014) 299.

⁵⁶ GDPR, arts 5(1)(b), 89(1). Cf LGPD, arts 7(IV), 16; Personal Data Protection Bill, No 373 of 2019, art 38 (India); Ley de Protección de Datos Personales No 29733 [2011], art 18(7) (Peru); PIPEDA, s 7(2) (Canada).

⁵⁷ Sartor and Lagioia (n 37) 81.

⁵⁸ Joseph A Cannataci and Jeanne Pia Mifsud Bonnici, 'The End of the Purpose-Specification Principle in Data Protection?' (2010) 24 International Review of Law, Computers and Technology 101; Forgó, Hänold, and Schütze (n 54).

⁵⁹ GDPR, art 5(1)(a); OECD, 'Collection Limitation' (n 18); Case C-524/06 *Heinz Huber v Bundesrepublik Deutschland* [2008] ECR I-09705, paras 54–62; LGPD, art 6(III); APPI, art 16.

possible design options and adopt the one that requires the least amount of data to achieve the desired outcome.

For AI, this means that measures should be taken in the deployment phase to minimise personal data required to derive a prediction. These can include, for instance, converting raw personal data into abstract numbers, or, related to storage limitation, hosting the model on the device used for the collection to avoid communicating some of the subject's data to an external host, as recommended by the British authority.⁶⁰

The principle of storage limitation imposes on the controller and the processor the obligation to store personally identifying data only for as long as strictly necessary for the specified purpose.⁶¹ For AI, this means that the data must be erased once the model is fully trained and once it is not strictly relevant anymore. For instance, if the model is meant to analyse data from the previous six months, data older than six months should be erased.

These principles are particularly relevant to ML processing. The main advantage of ML is its ability to analyse and draw patterns from enormous amounts of data in a time-efficient manner, which has proved fundamental in the advancement of scientific research and in the personalisation of services. To improve its analysis, ML requires that as much data as possible is used to train the model—a practice some call ‘data maximisation’.⁶² Moreover, data collected at first might become relevant again in future processing when coupled with other data sets.⁶³ To comply with data minimisation, controllers and processors would have to limit the collection and use of data to what is absolutely necessary. This would be difficult to achieve, potentially losing opportunities for future profit, research advancement, and improvements of personalised services, as future usefulness of data to make predictions is an insufficient reason to avoid the principles.⁶⁴

7 Accuracy in AI

The principle of accuracy, which has extended to numerous jurisdictions (eg Canada and Japan), is relevant to AI since it creates a right to a correct representation of oneself, so personal data must be correct and up to date.⁶⁵ Relatedly, controllers and processors must limit data processing while assessing data accuracy after a claim was made.⁶⁶ European courts have gone further and interpreted the

⁶⁰ ICO (n 30).

⁶¹ See eg California Privacy Rights Act, para 1798.100(c) (2020) (CPRA).

⁶² AIWP, 9.

⁶³ Norwegian Data Protection Authority, ‘Artificial Intelligence and Privacy’ (Report 2018) 18.

⁶⁴ See eg ICO (n 30); Sartor and Lagioia (n 37) 47–48.

⁶⁵ GDPR, art 5(1)(d); OECD, ‘Data Quality’ (n 18); Convention 108, art 5(3)(d); PIPEDA, s 4.6; APPI, art 19; LGPD, arts 6(V), 18(III).

⁶⁶ GDPR, art 18, Recital 67.

principle as supporting data subjects' right to erasure and rectification.⁶⁷ The right to erasure enables data subjects to demand that the controller and processor delete their personal data without undue delay. The right to rectification empowers data subjects to request the modification of inaccurate data or to supplement the information provided to correct wrongful outputs. Even if they are not always explicitly grounded on the accuracy principle, the rights to erasure and rectification, relevant to AI, are present outside of the GDPR, such as in Brazil and California (erasure)⁶⁸ and Australia (rectification).⁶⁹

Fulfilling these requests might adversely impact some of the benefits of ML. When a controller has to comply with a deletion request, they need to remove the data, which can worsen disparate outcomes.⁷⁰ So the principle of (data) accuracy can impact the principle of fairness. While removing data obtained for training purposes does not affect the system's function after the training phase has ended, doing so in the development phase would hinder system accuracy.⁷¹ Enforcing the right to rectification might not have a direct impact on the single individual during the training phase, but it might directly affect data subjects when it refers to the outputs of the model.⁷²

8 Security, Integrity, and Confidentiality in AI

Controllers and processors must also ensure that the data processing is done with appropriate security, integrity, and confidentiality.⁷³ They have the duty not to disclose private information to third parties and ensure that the system and storage of data are adequately protected from adversaries.⁷⁴ The principle also requires that data subjects and supervisory authorities are notified of data breaches,⁷⁵ which allows for timely countermeasures to avoid further harm.⁷⁶

⁶⁷ Eg Case C-362/14 *Maximillian Schrems v Data Protection Commissioner* ECLI:EU:C:2015:650, para 90.

⁶⁸ GDPR, art 17; cf LGPD, art 16; US, SB 1121, California Consumer Privacy Act (2018), para 1798.105 (CCPA).

⁶⁹ GDPR, art 16; Privacy Act 1998 (as amended 1 April 2022), s 14, principle 13 (Australia).

⁷⁰ Anya ER Prince and Daniel Schwarcz, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data' (2019) 105 Iowa Law Review 1257.

⁷¹ Christopher Kuner and others, 'Expanding the Artificial Intelligence-Data Protection Debate' (2018) 8 International Data Privacy Law 290.

⁷² Reuben Binns, 'Enabling Access, Erasure, and Rectification Rights in AI Systems' (Information Commissioner's Office, 15 October 2019) <<https://perma.cc/PKM9-DZK9>>.

⁷³ OECD, 'Security Safeguards' (n 18).

⁷⁴ See generally Charles P Pfleeger and Shari Lawrence Pfleeger, *Analyzing Computer Security: A Threat/Vulnerability/Countermeasure Approach* (Prentice Hall Professional 2012).

⁷⁵ GDPR, arts 33–34; cf LGPD, arts 6, 46; PIPEDA, s 10.1.

⁷⁶ Sébastien Gambs and others, *Privacy and Ethics: Understanding the Convergences and Tensions for the Responsible Development of Machine Learning* (Office of the Privacy Commissioner of Canada 2021).

AI can create new risks or worsen risks already associated with data processing.⁷⁷ Given that AI models are rarely fully developed and deployed in-house, their chain of design and use raises unique security concerns. This demands a holistic approach that assesses both the security of in-house developed AI and of externally sourced codes and frameworks. For example, related to data minimisation, developers might be asked to remove some features or to apply pseudonymisation techniques (which replace a subject's unique attribute for another reducing linkability) to increase the security of the training and testing data.⁷⁸ Developers should similarly routinely look for possible security vulnerabilities when developing a model using open-source code.⁷⁹

Security protections must address risks that are particular to ML models when a third party (such as a hacker) performs a privacy attack against a ML system to obtain personal data. Three risks stand out: model inversion, membership inference, and property inference. In model inversion attacks, an adversary infers information about individuals in a training data set that the attacker already had information about.⁸⁰ This type of attack could be used to reconstruct the training data set.⁸¹ Recalling the facial recognition example, an attacker might use a model inversion attack to recover and reconstruct the faces used to train the algorithm.⁸² In membership inference attacks, attackers deduce the presence of an individual in the model's training data set by looking at the confidence scores provided by the model's prediction.⁸³ When models are trained on a vulnerable or sensitive population (eg children or victims of sexual abuse), membership inference attacks present a significant risk for data subjects.⁸⁴ Property inference attacks infer properties about the training data set by exploiting properties that the model learned but do not work towards its task.⁸⁵ They can reveal sensitive information, such as the gender composition of the training data set.⁸⁶

⁷⁷ ICO (n 30); Sartor and Lagioia (n 37) 27–30.

⁷⁸ Article 29 Working Party, 'Opinion 05/2014 on Anonymisation Techniques' (10 April 2014) <<https://doi.org/10.1007/s10664-020-09830-x>>.

⁷⁹ Serena Elisa Ponta, Henrik Plate, and Antonino Sabetta, 'Detection, Assessment and Mitigation of Vulnerabilities in Open Source Dependencies' (2020) 25 Empirical Software Engineer 3175.

⁸⁰ Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, 'Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures' (22nd ACM SIGSAC Conference on Computer and Communications Security, New York, 2015) 1322–33.

⁸¹ Nicholas Carlini and others, 'The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks' (28th USENIX Security Symposium, 2019) 267–84.

⁸² See Pavana Prakash and others, 'Privacy Preserving Facial Recognition against Model Inversion Attacks' (IEEE Global Communications Conference, 2020).

⁸³ Reza Shokri and others, 'Membership Inference Attacks against Machine Learning Models' (IEEE Symposium on Security and Privacy, San Jose, 2017) 3–18.

⁸⁴ See eg Kashmir Hill and Gabriel JX Dance, 'Clearview's Facial Recognition App Is Identifying Child Victims of Abuse' *The New York Times* (10 February 2020) <<https://perma.cc/AJ3B-CYW8>>.

⁸⁵ Giuseppe Ateniese and others, 'Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers' (2015) 10 International Journal of Network Security 137–50.

⁸⁶ Mathias PM Parisot, Balazs Pejo, and Dayana Spagnuelo, *Property Inference Attacks, Convolutional Neural Networks, Model Complexity* (SECRYPT 2021).

Lastly, developers of ML models should consider the risk of adversarial examples to comply with the principle. These are modified examples aimed at being misclassified by the model. These can be concerning for data protection purposes when they impact data subjects' rights and freedoms. In the facial recognition example, an adversary could force the model to mistakenly recognise a manipulated picture as belonging to a different person.⁸⁷

9 Accountability in AI

The accountability principle is a 'meta-principle'⁸⁸ which renders data controllers responsible for demonstrating compliance with all the principles mentioned above and that of the processor.⁸⁹ The GDPR is particular in ensuring accountability through third-party auditing, the nomination of data protection officers,⁹⁰ and the conduction of data protection impact assessments—a development stemming from privacy impact assessments.⁹¹ Furthermore, accountability is twofold: controllers have both substantive and procedural obligations.⁹² First, being accountable implies that the controller takes proactive measures to ensure GDPR compliance by, for instance, implementing all necessary procedures to meet the principles by design and by default. Second, to demonstrate compliance with the data protection principles, the controller must record all aspects of the processing.

Because of the unpredictability of AI applications, Data Protection Impact Assessments (DPIA) have gained increased relevance.⁹³ In many jurisdictions, privacy and data protection laws demand the controller to undertake a DPIA,⁹⁴

⁸⁷ Mahmood Sharif and others, 'Accessorize to a Crime' (ACM SIGSAC Conference on Computer and Communications Security, New York, 2016) 1533.

⁸⁸ Reuben Binns, 'Data Protection Impact Assessments: A Meta-Regulatory Approach' (2017) 7 International Data Privacy Law 23, 27.

⁸⁹ Joseph Alhadeff, Brendan Van Alsenoy, and Jos Dumortier, 'The Accountability Principle in Data Protection Regulation: Origin, Development and Future Directions' in Daniel Guagnin and others (eds), *Managing Privacy through Accountability* (Palgrave Macmillan 2012) 49; OECD, 'Accountability Principle' (18); GDPR, arts 24, 28; PIPEDA, s 5.1; APEC, *Privacy Framework* (2005), 'Accountability Principle' <<https://perma.cc/5ZE-4JFD>>.

⁹⁰ GDPR, art 37; LGPD, art 41; PIPEDA, s 6, Sch 1 4.1.

⁹¹ See Roger Clarke, 'Privacy Impact Assessment: Its Origins and Development' (2009) 25 Computer Law and Security Review 123; Alessandro Mantelero, 'AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment' (2018) 34 Computer Law and Security Review 766. See also the chapter by Alessandro Ortalda and Paul De Hert in this volume.

⁹² GDPR, arts 5(2), 77, 82–83.

⁹³ Sonia K Katyal, 'Private Accountability in the Age of Artificial Intelligence' (2018) 66 UCLA Law Review 54, 115; Bryan Casey, Ashkon Farhangi, and Roland Vogl, 'Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise' (2019) 34 Berkeley Technology Law Journal 170–84; Mantelero (n 91); Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law and Technology Review 18, 77–80.

⁹⁴ GDPR, arts 35–36, Recitals 84, 89–95; LGPD, art 38; An Act to Modernize Legislative Provisions as regards the Protection of Personal Information, SQ 2021, c 25, s 3.3 (Quebec); Personal Information Protect Act, 2011 (as amended in 2020), art 33; cf CPRA, para 1798.185(a)(15)(B). See also Simon

which must specify (i) risks to individual rights posed by the data processing, and (ii) how such risks are mitigated according to the principles of necessity and proportionality. DPIAs play a fundamental role in the designing phase of the processing system by serving as a basis for the controller and the processor to identify the risks posed by their systems and, later, in the auditing phase by providing the means to demonstrate compliance.

Furthermore, controllers should continuously revise and reassess the system's DPIA throughout its life cycle.⁹⁵ Overall, DPIAs incentivise companies to take accountability for the risks (and their mitigation) that their system might pose.⁹⁶ Drawing back to the facial recognition example, if the controller decided to further process data for a compatible purpose, they would have to show proof of having performed a compatibility assessment, their decisional process and the safeguards put in place.⁹⁷

10 Conclusion

In this chapter, we outlined the main principles of privacy and data protection and their implications in the development and use of AI.

The growing use of AI has impacted societies worldwide. AI has transformed aspects as varied as medical innovations, personalisation of services, automated quality controls, and improvements in economic efficiency and competitiveness. But these innovations come with risks. Individuals' personal information should consequently be protected through a risk-based approach to prevent unreasonable uses of their data from these innovations adversely affecting their rights.

When deploying AI applications, data protection considerations might conflict with other interests. For instance, data minimisation can be counterproductive for statistical accuracy; transparency and explainability sometimes conflict with commercial secrecy; and purpose and storage limitation can obstruct innovation. Controllers should take such trade-offs into their considerations over the risks

Reader, 'Guidance on AI and Data Protection' (*Information Commissioner's Office*, 30 July 2020) <<https://ico.org.uk/about-the-ico/media-centre/ai-blog-data-protection-impact-assessments-and-ai/>>.

⁹⁵ Article 29 Data Protection Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679' (2017) 14.

⁹⁶ Margot E Kaminski and Gianclaudio Malgieri, 'Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations' (2021) 11(2) International Data Privacy Law 125–44.

⁹⁷ GDPR, art 6(4).

posed by the AI model being used. Controllers should continuously assess such risks and identify reasonably available technical mitigation tools.⁹⁸ Privacy and data protection considerations lie at the core of AI development. For AI to be compliant with the law and with the right to privacy, these considerations cannot be disregarded.

⁹⁸ GDPR, arts 5(2), 24.

The Rights to Privacy and Data Protection and Facial Recognition Technology in the Global North

Natalia Menéndez González

1 The Right to Privacy: A Corollary for Other Fundamental Rights

The objective of this chapter is to offer an interpretation of the rights to privacy in general, and data protection in particular, as a corollary, an umbrella to other fundamental rights. I consider this interpretation particularly relevant within the law and technology field, especially when dealing with artificial intelligence (AI). This is because the impact of AI systems, apart from transversal, intertwines diverse fundamental rights. Such an impact affects many levels, tracing networks between one fundamental right and the other. For instance, the use of facial recognition technology (FRT) has been flagged as potentially risky since it allows for the mass processing of biometric data. Additionally, mass surveillance performed by FRT on public demonstrations can also entail a deterrence effect on the participants, thus affecting their right to freedom of assembly and association, as demonstrated, for instance, in India.¹

If a fundamental right that acts as a backbone to the others is to be named, as far as AI is concerned, the right to privacy and the right to data protection are the main candidates. This is because AI systems rely on vast amounts of data to produce their outputs. This data is not always personal data, but taking into account the current level of (Big) data collection and processing, and the fact that it is relatively easy from two or three pieces of information to de-anonymise an individual,² it can be said that (personal) data are the predominant food for AI systems.

Further, this intertwining of the rights to privacy and data protection has also been pointed out by several stakeholders that have discussed how the impact of

¹ Jay Mazoomdaar, 'Delhi Police Film Protests, Run its Images Through Face Recognition Software to Screen Crowd' *The Indian Express* (28 December 2019) <<https://indianexpress.com/article/india/police-film-protests-run-its-images-through-face-recognition-software-to-screen-crowd-6188246/>>.

² M Kearns and A Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (OUP 2020).

AI technologies on the rights to privacy and data protection had a collateral effect on other fundamental rights. Such an impact should not be overlooked as redress ought to reach all fundamental rights affected, no matter the extent.³

To successfully convey the task of this chapter, being to analyse the impact of AI on the rights to privacy and data protection understood as a corollary for other fundamental rights, the example of FRT is chosen. This AI-empowered technology is a perfect case study to show the impact of AI systems on fundamental rights in general but on the rights to privacy and data protection in particular. The argument expounded in this chapter is limited to the Global North. Thus, I will mainly refer to Convention 108+,⁴ the European Union's (EU) General Data Protection Regulation (GDPR),⁵ and the Illinois Biometric Information Privacy Act (BIPA) in the United States (US) to address the regulatory issues of privacy and data protection for FRT.

To better understand the actual reach of FRT, section 2 conducts a brief technical introduction to the technology. Section 3 explains the nature of AI technologies as (personal) data-driven ones, emphasising the important role of the rights to privacy and data protection. Section 4 exemplifies, using FRT, how such role extends to other fundamental rights, thereby presenting the justification of the rights to privacy and data protection as a corollary of other fundamental rights.

2 Facial Recognition Technology: An AI-Empowered Big Brother

Facial recognition technology (FRT) compares two or more facial images of an individual to determine the likelihood that such facial images belong to the same individual. Before the introduction of machine learning (ML) techniques to FRT, the accuracy of such systems was harshly questioned. Facial images of people in movement, wearing glasses, or other items on their heads were not suitable for pre-ML-empowered FRT. It only worked (and not within 100 per cent of the cases) with ID-style facial portraits under certain lighting conditions.⁶ These 'laboratory'

³ For a possible limit to this approach, see the chapter on environmental rights by Alberto Quintavalla in this volume.

⁴ Council of Europe's Convention 108+ for the Protection of Individuals with regard to the Processing of Personal Data (ETS No 181)

⁵ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (General Data Protection Regulation/GDPR).

⁶ Kelly A Gates, *Our Biometric Future* (NYU Press 2011).

conditions made the technology practically useless since the proposed uses for FRT such as checking at borders, suspect or missing people identification, or monitoring workers' attendance do not often occur within the above-mentioned conditions.

However, the introduction of ML techniques radically changed that scenario. As a result, FRT is capable of recognising people in movement, with glasses or hats, not looking directly at the camera, and even wearing facial masks. Lately, the introduction of ML techniques has even expedited the face detection and features extraction steps streamlining the recognition process.

The addition of AI has also impacted FRT by allowing such systems to perform a third function, namely categorisation. Together with the identification (one-to-many) and verification/authentication (one-to-one) functions, there is a wide debate about what has been called the categorisation function of FRT. Such a function consists of allocating certain features to an individual based on the analysis of their facial picture. There is no wide consensus on whether face categorisation can be considered facial recognition or not. On the one hand, bodies such as the Article 29 Working Party (the European Data Protection Board's predecessor) considered categorisation a facial recognition function. On the other, from a technical point of view,⁷ recognition only covers identification and verification. Such a debate has not prevented the proliferation of categorisation applications as the ones claiming to infer sexual orientation,⁸ political beliefs,⁹ or criminality¹⁰ from facial analysis. However, the facial analysis research stream has been widely criticised based on claims of its lack of supporting scientific evidence.¹¹

This brief explanation of what FRT is, how it works, and what the addition of AI entailed to it already hints at some of the risks that such a technology might pose for fundamental rights with the rights to privacy and data protection within the frontline. Therefore, section 3 will directly analyse the impact of FRT on such rights.

⁷ See the ISO/IEC 2387-37:2017 definition of 'biometric recognition' under term 37-01-03 on the harmonised biometric vocabulary.

⁸ Yilun Wang and Michal Kosinski, 'Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images' (2018) 114(2) *Journal of Personal and Social Psychology* 246.

⁹ Michal Kosinski, 'Facial Recognition Technology Can Expose Political Orientation from Naturalistic Facial Images' (2021) 11(1) *Scientific Reports* 100.

¹⁰ Xiaolin Wu and Xi Zhang, 'Automated Inference on Criminality Using Face Images' (2016) arXiv:1611.04135v2 [cs.CV] 21 Nov 2016; and Xiaolin Wu and Xi Zhang, 'Responses to Critiques on Machine Learning of Criminality Perceptions' (2016) arXiv:1611.04135v3 [cs.CV] 26 May 2017.

¹¹ Luke Stark and Jevon Hutson, 'Physiognomic Artificial Intelligence' (2022) 32 *Fordham Intellectual Property, Media and Entertainment Law Journal* 922.

3 AI as Personal Data-Driven and the Rights to Privacy and Data Protection

As mentioned within section 1, the increasing presence of Big Data (bases) has been the perfect companion to and one of the main facilitating factors behind the current build-up of AI. From recidivism scoring, social funds allocation, and CV screening, to FRT, there is a myriad of AI applications that are fuelled by personal data.

Personal data has been defined by Convention 108+ as ‘any information relating to an identified or identifiable individual (“data subject”):¹² The GDPR goes one step further and adds to such definition that:

[A]n identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.¹³

If we move to FRT, facial images—when technically processed to identify a person—are considered biometric data, and this is explicitly stated within article 4(14) of the GDPR. Further, ‘biometric data for the purpose of uniquely identifying a natural person’ are deemed sensitive data according to article 9(1) of the GDPR,¹⁴ and a very similar formulation is adopted by Convention 108+, deeming biometric data ‘special categories of data’.¹⁵ Finally, under the BIPA, ‘biometric information’ is any information—regardless of how it is captured, converted, stored, or shared—based on an individual’s biometric identifier used to identify an individual.¹⁶

Therefore, facial images processed by FRT are subject to data protection regulation. As stated by the Information Commissioner’s Office (ICO)¹⁷, ‘[s]ensitive processing occurs irrespective of whether that image yields a match to a person on a watchlist or the biometric data of unmatched persons is subsequently deleted within a short space of time’.¹⁸

Sensitive data processing (and, therefore, FRT) is, in principle, prohibited by article 9(1) of the GDPR. Convention 108+ only allows for the processing of such data ‘where appropriate safeguards are enshrined in law, complementing those of this

¹² Convention 108+, art 2.a.

¹³ GDPR, art 4.1.

¹⁴ See also GDPR, Recital 51.

¹⁵ Convention 108+, art 6.1.

¹⁶ Biometric Information Privacy Act 740 ILCS 14, s 10 (BIPA).

¹⁷ The United Kingdom’s data protection authority

¹⁸ Information Commissioner’s Office (ICO), ‘The Use of Live Facial Recognition Technology by Law Enforcement in Public Places (2019/01)’ (ICO, 2019) <<https://ico.org.uk/media/about-the-ico/documents/2616184/live-frt-law-enforcement-opinion-20191031.pdf>>.

Convention.¹⁹ Hence, it contemplates the corollary nature of the rights to privacy and data protection by establishing that '[s]uch safeguards shall guard against the risks that the processing of sensitive data may present for the interests, rights and fundamental freedoms of the data subject, notably a risk of discrimination'.²⁰

Article 9(2) of the GDPR establishes a series of exceptions to the general prohibition of biometric data processing. This list of exceptions sketches the only situations where biometric data processing (and, therefore, FRT) is allowed under EU law. Accordingly, it also gives us an impression of the level of restriction that such processing is subjected to due to the sensitivity of facial images. This is because, in the case of a biometric data breach, facial images are not interchangeable, replaceable, or erasable such as passwords or credit cards. Faces cannot be changed, and this stands—notwithstanding some current movements which encourage citizens to 'dress up' their faces to avoid FRT.²¹ This entails a great loss for the affected data subjects and deeply infringes their fundamental rights to privacy and data protection.

As facial images are the 'food' for FRT and they are personal data as long as they fit within the definitions discussed above, FRT and also many (other) AI technologies would be personal data-driven. Therefore, the rights to privacy and data protection grant a crucial position within the overall body of fundamental rights since there can be no FRT or many other AI systems without personal data. This does not mean that other fundamental rights (some of which will be discussed within section 4) are less important or should not be taken into account if privacy and data protection are respected. What it means is that privacy and data protection act as a corollary, an umbrella right. Without respect for such rights, AI systems in general, and FRT in particular, do not comply with fundamental rights by design and by default (using the GDPR's terminology).²² In that event, not even the use of such technologies should be allowed or even tested in the first place.

Further, as will be seen in section 4, violations of fundamental rights other than privacy and data protection by AI systems sometimes connect with privacy and data protection considerations. For instance, the right to freedom of assembly may be violated in conjunction with the rights to privacy and data protection in instances where location data from a person attending a demonstration is not protected.

Moreover, violations of privacy and data protection rights can affect other rights, since consent for being subjected to an AI system does not reach all the possible

¹⁹ ibid.

²⁰ Convention 108+, art 6.2.

²¹ Tom Simonite, 'How to Thwart Facial Recognition and Other Surveillance' (*Wired*, 22 September 2020) <www.wired.com/story/how-to-thwart-facial-recognition-other-surveillance/>; and Aaron Holmes, 'These Clothes Use Outlandish Designs to Trick Facial Recognition Software into Thinking You're Not Human' (*Business Insider*, 5 June 2020) <www.businessinsider.com/clothes-accessories-that-outsmart-facial-recognition-tech-2019-10>.

²² GDPR, art 25.

and (sometimes unsuspected) future uses or repurposes of that system, with unpredictable consequences for fundamental rights in general.

Accordingly, on the issue of consent, we discern the following requirements. The BIPA prohibits private companies from collecting biometric information unless they (1) inform the person in writing of (a) what data is being collected or stored, (b) the specific purpose and length of time for which the data will be collected, stored, and used and (2) obtain the person's written consent.²³ In the same line, regarding the lawful grounds for data processing, article 4(11) of the GDPR determines that consent should be a freely expressed, explicit, informed, and unequivocal expression of the data subject's intentions by which they affirmatively signal acceptance of the processing of their personal data. This definition is complemented by Recital 32 of the GDPR, which states that consent should apply to all processing actions performed for the same or similar reasons. When there are numerous reasons for the processing, consent should be provided for all of them. Consent is a fundamental—and often misused—piece within any data protection regime.

Academics are divided regarding the real scope of consent. Some authors argue that consent is irrelevant—generally, the consent does not depend on actual information about the processing or consent is not truly a real option. That is, the issue of personal data processing today (especially in FRT and AI contexts) is so complex that most people lack the expertise to grasp it and are unable to foresee the risks involved.²⁴

Moreover, a denial of consent might imply that services that are—socially, economically, or otherwise—essential may be out of reach (or limited) for those rejecting to be subjected to FRT, affecting the fundamental rights to human dignity and non-discrimination. In the same line, according to Recital 43 of the GDPR to ensure that consent is freely given, it should not be used as a valid legal basis for processing personal data when there is a clear imbalance between the data subject and the controller, particularly where the controller is a public authority, and it is thus unlikely that consent was freely given in all the circumstances. This is the case, according to the Swedish DPA and the *Tribunal administratif de Marseille*, when consenting to the use of FRT within schools.²⁵ The DPA stated that students

²³ BIPA, s 15.

²⁴ Evan Selinger and Woodrow Hartzog, 'The Inconsentability of Facial Surveillance' (2020) 66 Loyola Law Review 101; Stefan Schiffner and others, 'Towards a Roadmap for Privacy Technologies and the General Data Protection Regulation: A Transatlantic Initiative' (Annual Privacy Forum, Barcelona, 2018); Giovanni Sartor and Francesca Lagioia, *The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence* (European Parliamentary Research Service 2020); and Genia Kostka, Léa Steinacker, and Miriam Meckel, 'Between Security and Convenience: Facial Recognition Technology in the Eyes of Citizens in China, Germany, the United Kingdom, and the United States' (2021) 30(6) Public Understanding of Science 671.

²⁵ Swedish Data Protection Authority, 'Skellefteå Municipality, Secondary Education Board Supervision Pursuant to the General Data Protection Regulation (EU) 2016/679: Facial Recognition Used to Monitor the Attendance of Students' Ref No DI-2019-2221 (2019); and Tribunal Administratif de Marseille, 'La Quadrature Du Net et autres' No 1901249 (27 February 2020).

'depended' on schools for questions related to their academic future and therefore, their consent (or their guardian's) might not be valid.²⁶ Similarly, this can be claimed to be the case of employees being subjected to their employers' desires or power dynamics.²⁷

Other authors establish that consent does not cover the potential—and often undefined—use of data, even when such use is socially advantageous.²⁸ This is precisely the case pointed to by academics of an apparent incompatibility between Big Data and consent.²⁹ Using Big Data to train AI systems and allow them to make inferences might be contradictory to the lawfulness of processing, in the sense that some of the purposes of the training phase within the learning process of an AI system could not be anticipated. The AI-empowered system could end up making inferences that the data processor might not have anticipated and, therefore, collected consent for. Moreover, some research has pointed out that, in the case of FRT, consent might not be free and informed (because both the uses and functioning of the technology are uncertain to some extent, due to its innovative and 'black box' nature) and, therefore, the processing might be considered unlawful.³⁰ The different functions that FRT may perform have to be also taken into account. In a certain scenario, the data subject might end up 'swamped' with consent requests.

Along similar lines, several authors have spotted a risk to privacy when biometric data, in general, are used for secondary purposes not compatible with the ones for which the data were initially collected. They place special emphasis on cases where third parties with access to facial images, such as law enforcement agents, merge them along with other data without consent from the data subject.³¹ This might present problems, for instance, in cases where FRT is built on top of a different system, such as CCTV or a thermal scanner. Again, due to the innovative nature of the technology, people may not possess enough knowledge to understand the true impact of what they are consenting to.³²

Regarding permission for biometric data processing for reasons of substantial public interest, from the perspective of the fundamental right to security, this clause leaves room for uncertainty since this justification has not been defined

²⁶ Swedish Data Protection Authority (n 25).

²⁷ Slovenian Data Protection Authority, 'List of Consents to Workplace Body Temperature Measurement' (No 07121-1/2020/1774, 2020).

²⁸ Fred H Cate, Peter Cullen, and Viktor Mayer-Schonberger, *Data Protection Principles for the 21st Century: Revising the 1980 OECD Guidelines* (Microsoft 2013).

²⁹ Rosario Girasa, *Intelligence as a Disruptive Technology* (Palgrave Macmillan 2020); Schiffner (n 24); Sartor and Lagioia (n 24); Sandra Wachter, 'Data Protection in the Age of Big Data' (2019) 2(1) *Nature Electronics* 6.

³⁰ Selinger and Hartzog (n 24); Sartor and Lagioia (n 24); Schiffner (n 24).

³¹ Ann Cavoukian, *Privacy and Biometrics* (Information and Privacy Commissioner 1999); and Ioannis Iglezakis, 'EU Data Protection Legislation and Case-Law with Regard to Biometric Applications' (18 June 2013) <<https://ssrn.com/abstract=2281108>> or <<http://dx.doi.org/10.2139/ssrn.2281108>>.

³² Selinger and Hartzog (n 24); Schiffner (n 24); Sartor and Lagioia (n 24); and Kostka (n 24).

within data protection regulation and, therefore, it might have multiple interpretations.³³ For instance, Recital 45 of the GDPR leaves the task to lay down such a public interest concept in the hands of EU or member states law leading to a fragmented picture within the EU regulatory landscape. This might also open the door to a 'blanket justification' of FRT deployment if not carefully assessed.

On the other hand, the BIPA, in addition to its notice and consent requirement, prohibits any company from selling or otherwise profiting from consumers' biometric information.³⁴ Therefore, it follows that a reliable and predictable legal basis must be established by the controller before starting the processing of personal data by FRT. However, disrespect for these measures might not only impact the rights to privacy and data protection but other rights. Section 4 will unveil how the impact of AI and FRT on the rights to privacy and data protection might also affect other fundamental rights.

4 The Impact of Privacy Violations by AI on Other Fundamental Rights

No matter how big the impact of FRT on the rights to privacy and data protection, it is often just a single facet within a much broader impact on other fundamental rights. Driven by the mass processing of biometric data, FRT most likely impacts the above-mentioned rights. However, FRT's use puts such data processing to the service of other activities, mainly (but not only) related to surveillance. This section will explore the influence of FRT's usage on other fundamental rights.

4.1 Human Dignity

Any infringement of the rights to privacy and data protection will subsequently impact the dignity of the subject. This is the argument employed by both the European Data Protection Supervisor (EDPS) and the European Court of Human Rights (ECtHR) when referring to mass surveillance.

In the case *S and Marper v the United Kingdom*, the ECtHR established that 'blanket and indiscriminate' retention of biometric data interferes with the right to privacy and may also have a stigmatising effect, treating persons presumed innocent alike with criminals.³⁵ Along similar lines, the EDPS has underlined that the

³³ Information Commissioner's Officer (ICO), 'What Are the Substantial Public Interest Conditions?' < <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-categories-data/what-are-the-substantial-public-interest-conditions/>>.

³⁴ BIPA, s 15.

³⁵ *S and Marper v the United Kingdom* App nos 30562/04 and 30566/04 (ECtHR, 4 December 2008), para 122.

commodification and objectification of people's faces, especially by algorithms and for the benefit of state surveillance, to be used on a large-scale scheme, contravene dignity.

Such commodification and objectification of people's physical features has found its perfect spot within the current Big Data environment. It is now argued that human beings have been reduced to pieces of data—starting from their biometric features and continuing through their behaviour or search habits on the Internet—and, as a result, every human being can be diminished to a bag set of data.³⁶ This data may be exploited by companies for profit—especially Big Tech.³⁷ Further, extracting and processing facial images for surveillance purposes enhances a 'datafication' of our facial features.³⁸ This collides with our traditional vision of human faces as a mirror of the soul and a highly important proxy of a person's self.³⁹ Therefore, treating faces as another element for identification, like our ID number, will perpetuate this depersonalising vision, and thus contravene human dignity.

4.2 Security

Scholarship has extensively analysed the balance between the right to security and the operationalisation of such a right, for example, in implementing surveillance measures to enforce or enhance it.⁴⁰ The weighing operation between the right to security and the rights to privacy and data protection is in line with the principle of proportionality as understood by the constitutional law doctrine ('rights balancing'). However, the entrance of modern FRT systems has entailed another turn of the screw. As previously explained, modern FRT identifies people much more efficiently than humans can do.⁴¹ Also, the risks of using FRT are much higher than using human-empowered recognition. Consequently, the use of FRT must be subject to the principle of proportionality to prevent such interferences the rights are precisely trying to safeguard.

³⁶ John Cheney-Lippold, *We are Data Algorithms and the Making of Our Digital Selves* (NYU Press 2017).

³⁷ Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (Profile 2019).

³⁸ Cheney-Lippold (n 36).

³⁹ Jonathan Cole, 'Empathy Needs a Face' (2001) 8(5–6) *Journal of Conscious Studies* 51.

⁴⁰ Miriam Dornan, 'Security vs Liberty? Is There a Trade Off' (*E-International Relations Publishing*, 23 June 2011) <www.e-ir.info/2011/06/23/security-vs-liberty-is-there-a-trade-off/>; Vincenzo Pavone, Elvira Santiago Gomez, and David-Olivier Jaquet-Chifelle, 'A Systemic Approach to Security: Beyond the Tradeoff Between Security and Liberty' (2016) 12(4) *Democracy and Security* 225; and Adrian Vermeule, 'Security and Liberty: Critiques of the Trade-off Thesis' in David Jenkins and others (eds), *The Long Decade: How 9/11 Changed the Law* (OUP 2014) 31.

⁴¹ P Jonathon Phillips and Alice J O'Toole, 'Comparison of Human and Computer Performance Across Face Recognition Experiments' (2014) 32(1) *Image and Vision Computing* 74.

The principle of proportionality, as articulated in case law and public law doctrine, establishes guidelines for justifying interference with people's fundamental rights and freedoms. It is a general principle of law that has transcended its German origins and travelled throughout the world to become a universally acknowledged law principle. The European Court of Justice (ECJ) has defined the principle within its *Gebhard* judgment of 1995,⁴² stating that:

[M]easures liable to hinder or make less attractive the exercise of fundamental freedoms guaranteed by the Treaty must fulfil four conditions: they must be applied in a non-discriminatory manner; they must be justified by imperative requirements in the general interest; they must be suitable for securing the attainment of the objective which they pursue; and they must not go beyond what is necessary in order to attain it.⁴³

These criteria present in *Gebhard* might also be applied to FRT. First, various—and, moreover, intersectional—inaccuracies have been reported, notably including both gender and ethnicity discriminatory outcomes.⁴⁴ Moving on to the second prong—‘justification by imperative requirements in the general interest’—this might be the case when FRT is, for instance, used by law enforcement or for health reasons. However, no extensive data are available to back up to what extent FRT has helped law enforcement operations or reinforced health measures within the context of the COVID-19 pandemic, for instance.⁴⁵ Similar arguments might be applied to the suitability and necessity requirements. FRT has not proved, up to this moment, a significant technological advantage when confronted against other technologies for identification or verification/authentication that justify surpassing the proportionality and necessity threshold.⁴⁶ In this line, the Dutch DPA—apropos the use

⁴² Case C-55/94 *Reinhard Gebhard v Consiglio dell'Ordine degli Avvocati e Procuratori di Milano* [1995] ECR I-4186.

⁴³ *ibid.*

⁴⁴ Joy Buolamwini and Timnit Gebru, ‘Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification’ (1st Conference on Fairness, Accountability and Transparency, New York City, 2018).

⁴⁵ See David Spiegelhalter and Kevin Mcconway, ‘Live Facial Recognition: How Good Is It Really? We Need Clarity About the Statistics’ (Winton Centre, 10 February 2020) <<https://medium.com/wintoncentre/live-facial-recognition-how-good-is-it-really-we-need-clarity-about-the-statistics-5140bd3c427d>>. The only data publicly available at the moment on the use of FRT by LE that can shed some light about its effectiveness are the following: INTERPOL, ‘Fact Sheet Facial Recognition’ (COM/FS/2020-03/FS-04, 2020) <www.interpol.int/How-we-work/Forensics/Facial-Recognition>; ‘Almost 1,500 terrorists, criminals, fugitives, persons of interest or missing persons have been identified since the launch of INTERPOL’s facial recognition system at the end of 2016’; and Eldar Haber, ‘Racial Recognition’ (2021) 43(1) Cardozo Law Review 71: ‘In 2019, the Facial Identification Section received 9,850 requests for comparison and identified 2,510 possible matches, including possible matches in 68 murders, 66 rapes, 277 felony assaults, 386 robberies, and 525 grand larcenies’.

⁴⁶ Luana Pascu, ‘EU Regulators Warns Facial Recognition in Law Enforcement, Commercial Settings May Be Illegal’ (*Biometric Update*, 11 June 2020) <www.biometricupdate.com/202006/eu-regulators-warns-facial-recognition-in-law-enforcement-commercial-settings-may-be-illegal>; ‘However, the security of a supermarket is not so important that biometric data can be processed for this’.

of FRT in sports competitions—said that if FRT only helps to a limited extent and does not reveal additional benefits, its use is disproportional.⁴⁷

In conclusion, the use of FRT might help to enhance the rights to security or liberty. However, if the fine line between a restrictive interpretation of public interest and mass surveillance is crossed, the same technology deployed to protect such rights might be the one that threatens them.

4.3 Freedom of Assembly and Association

The deployment of such a mass surveillance apparatus might also inhibit the rights to freedom of assembly and association by deterring people to exercise such rights for fear to be targeted or punished in some way.⁴⁸ There have been numerous cases all around the world where FRT has been deployed to monitor protesters.⁴⁹

This situation reached a point where Russian activist Alyona Popova and the former Deputy Minister of Energy lodged a complaint to the ECtHR. The complaint, the first one with the ECtHR concerning FRT, was based on the use of such technology by the Moscow authorities at an authorised rally in support of those detained and prosecuted for participating in peaceful protests about the barring of independent candidates from the Moscow municipal assembly elections.⁵⁰ According to the claimants, all protesters in September had to go through metal detectors with CCTV cameras positioned at eye level. The Moscow administration had revealed intentions to employ FRT at big public gatherings prior to the demonstrations. According to the petitioners, this is the first instance of the Moscow authorities employing FRT to acquire data on demonstrators.

The use of FRT to monitor protesters might prevent people to attend such demonstrations to avoid being identified. This chilling effect would undermine the right to freedom of assembly and association since people will think twice or even stop expressing their beliefs at public displays out of a fear of being tracked and recognised by FRT.

⁴⁷ Autoriteit Persoonsgegevens (Dutch Data Protection Authority/DPA), ‘Vragen over inzet gezichtsherkenning [Questions about the use of facial recognition]’ (z2003-1529, 3 February 2004).

⁴⁸ See the chapter by Margaret Warthon in this volume.

⁴⁹ Evan Selinger and Albert Fox Cahn, ‘Did You Protest Recently? Your Face Might Be in a Database’ *The Guardian* (17 July 2020) <www.theguardian.com/commentisfree/2020/jul/17/protest-black-lives-matter-database>.

⁵⁰ Anastasiia Kruope, ‘Moscow’s Use of Facial Recognition Technology Challenged’ (*Human Rights Watch*, 8 July 2020) <www.hrw.org/news/2020/07/08/moscow-s-use-facial-recognition-technology-challenged>.

4.4 Equality Before the Law and Non-Discrimination

Algorithmic discrimination has been the hobbyhorse of AI in general, and FRT in particular. This problem was brought to the mainstream light by Buolamwini et al who demonstrated that FRT struggled to identify Black people—especially women—and that the rate of false positives was disproportionately high compared with white people—especially men.⁵¹ The consequences of such inaccuracies deeply affect fundamental rights, with the wrongful arrest of Black men due to incorrect identification by FRT just one example.⁵² In the same vein, scholarship is starting to point out the discriminatory potential that FRT in general, and automatic gender recognition in particular, entails for people with trans- or non-binary gender identities.⁵³

But not only the technology but also its use has been instrumentalised to discriminate against minorities. In China, FRT in its categorisation and identification functions is used to monitor and oppress people from the Uyghur ethnic minority within the province of Xinjiang.⁵⁴

Along with the potential to become a tool for mass surveillance, the wide presence of algorithmic bias within the technology—plus the potential risk for automation bias—will vulnerable the fairness principle of data processing but also the fundamental right to equality before the law and non-discrimination .

4.5 The Rights of the Child

The rights of the child require special and reinforced protection due to the vulnerability of the subject. For instance, article 8 of the GDPR specifically regulates the conditions governing children's consent when accessing information society services.⁵⁵ In this line, article 15(2)(e) of Convention 108+ establishes that 'specific attention shall be given to the data protection rights of children'. When addressing the issue of blanket retention of biometric data for law enforcement purposes of people who have not been convicted of a crime, the ECtHR emphasised in *S and Marper v the United Kingdom* that this may be especially harmful in the case of

⁵¹ Buolamwini and Gebru (n 44).

⁵² Khari Johnson, 'How Wrongful Arrests Based on AI Derailed 3 Men's Lives' (*Wired*, 7 March 2022) <www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives/>.

⁵³ Os Keyes, 'The Misgendering Machines: Trans/HCI implications of Automatic Gender Recognition' (2018) 2 Proceedings of the ACM on Human-Computer Interaction 1; and Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker, 'How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services' (2019) 3 Proceedings of the ACM on Human-Computer Interaction 1. See also the chapter by Masuma Shahid in this volume.

⁵⁴ Jane Wakefield, 'AI Emotion-Detection Software Tested on Uyghurs' (*BBC News*, 26 May 2021) <<https://www.bbc.com/news/technology-57101248>>; and Richard Van Noorden, 'The Ethical Questions that Haunt Facial-Recognition Research' (2020) 587(7834) *Nature* 354.

⁵⁵ See also GDPR, Recitals 38, 58.

children, given their unique situation and the importance of their development and integration into society.⁵⁶ Children are particularly challenging for FRT since the precision of a biometric match decreases as they are constantly changing physically. According to the Fundamental Rights Agency, when face pictures that were taken at a young age are compared with those taken more than five years later, the likelihood of a mismatch increases.⁵⁷ In general, studies show that the accuracy of FRT is substantially worse for children under the age of thirteen. Further, current FRT only guarantee a solid match if the child was at least six years old when the biometric face image was collected and the match occurred within a five-year time frame.⁵⁸ Consequently, software studies show that photos of younger individuals result in far more false negatives than images of older people.⁵⁹

This might collide with the proposal to reduce the age of children for biometric data collection from fourteen to six years old in the context of Eurodac, the European system for the comparison of fingerprints of asylum applicants.⁶⁰ This same proposal contemplates that the EU Agency for the Operational Management of Large-Scale IT Systems (eu-LISA) would perform a study on the technical feasibility of integrating face recognition software based on facial picture data held in Eurodac within three years after adoption. While the modified Eurodac plan includes the need to identify children who are missing (abducted), victims of a crime, or may have been separated from their families, the use of FRT might help children to preserve or re-establish their identity according to the mandate based on article 8 of the Convention on the Rights of the Child (CRC) (duty to respect the right of the child to preserve his or her identity, including nationality, name, and family relations). Thus, we might resort again to the principle of proportionality to wonder whether solving such cases justifies the massive collection of facial images of young migrants and asylum seekers.

⁵⁶ *S and Marper v the United Kingdom* (n 35) paras 124–25.

⁵⁷ EU Fundamental Rights Agency, ‘Under Watchful Eyes: Biometrics, EU IT Systems and Fundamental Rights’ (2018) 109 <https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-bio-metrics-fundamental-rights-eu_en.pdf>.

⁵⁸ Sakshi Sahni, ‘SURVEY: Techniques for Aging Problems in Face Recognition’ (2014) 4(2) MIT International Journal of Computer Science and Information Technology 1; Narayanan Ramanathan, Rama Chellappa, and Soma Biswas, ‘Computational Methods for Modeling Facial Aging: A Survey’ (2009) 20(3) Journal of Visual Languages and Computing 131; and Javier Galbally Herrero and others, *Study on Face Identification Technology for its Implementation in the Schengen Information System* (Publications Office of the European Union 2019).

⁵⁹ Patrick Grother and others, *Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification* (NIST 2022); Nisha Srinivas and others, ‘Face Recognition Algorithm Bias: Performance Differences on Images of Children and Adults’ (IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, 2019).

⁶⁰ Bianca-Ioana Marcu, ‘Eurodac: Biometrics, Facial Recognition, and the Fundamental Rights of Minors’ (*European Law Blog*, 29 April 2021) <<https://europeanlawblog.eu/2021/04/29/eurodac-bio-metrics-facial-recognition-and-the-fundamental-rights-of-minors/>>.

Furthermore, the plan to gather children's face pictures for future FRT may give them the impression that they are continuously being monitored, thus deterring their free development and affecting their fundamental rights.

5 Conclusion

AI applications in general, and FRT in particular, have a great impact on the rights to privacy and data protection since such technologies are fed by vast amounts of (often personal) data. However, such an impact on the rights to privacy and data protection might transcend to other rights such as (1) human dignity, (2) security, (3) freedom of assembly and association, (4) equality before the law and non-discrimination, and (5) the rights of the child. This grants, to the rights to privacy and data protection, the status of corollary of other fundamental rights; violations of the rights to privacy and data protection by FRT might end up impacting other rights, and violations by FRT of other rights might have a privacy and data protection component..

Privacy, Political Participation, and Dissent

Facial Recognition Technologies and the Risk of Digital Authoritarianism in the Global South

Malcolm Katrak and Ishita Chakrabarty

1 Introduction

Facial recognition technologies (FRTs) are at the forefront of surveillance technologies fuelled by artificial intelligence (AI) that are changing the nature of law enforcement across the globe.¹ FRTs constitute a subdomain of computer vision, which, in turn, is a domain of AI that allows AI systems to ‘extract information, build models, and apply multi-dimensional data in a broad range of computer-generated activities’.² FRTs essentially entail the capturing of an image, the generation of a digital code known as a ‘faceprint’ using an algorithm, and the use of an algorithm to compare the captured image to a database of stored images.³ According to the Carnegie Endowment for International Peace, 64 out of 176 countries that were the subject of a study were found to be utilising FRTs for surveillance purposes,⁴ signifying the pervasive reach of FRTs in this regard. Significantly, the AI Global Surveillance Index developed by Carnegie, which classifies regimes as closed autocracies, electoral autocracies, electoral democracies, and liberal democracies, found the use of FRTs to be prevalent amongst regimes falling into all four categories.⁵ The burgeoning use of FRTs entails the breaking down of the binary between authoritarian regimes and state surveillance, with its relative ubiquity in both liberal democratic and authoritarian states raising concerns relating to the erosion of the right to privacy.⁶ The usage of FRTs in public spaces raises concerns

¹ Andrew Ferguson, ‘Facial Recognition and the Fourth Amendment’ (2021) 105 Minnesota Law Review 1105, 1107.

² Rosario Girasa, *Artificial Intelligence as a Disruptive Technology* (Palgrave Macmillan 2020) 17.

³ United States Government Accountability Office, ‘Facial Recognition Technology: Commercial Uses, Privacy Issues, and Applicable Federal Law’ (GAO-15-621) (July 2015) <www.gao.gov/assets/gao-15-621.pdf>.

⁴ Steven Feldstein, ‘The Global Expansion of AI Surveillance’ (2019) Carnegie Endowment for International Peace Working Paper, 19 <https://carnegieendowment.org/files/WP-Feldstein-AISurveillance_final1.pdf>.

⁵ ibid 25–28.

⁶ Peter Königs, ‘Government Surveillance, Privacy, and Legitimacy’ (2022) 35(8) Philosophy and Technology 8, 9.

not only relating to privacy but also democratic participation itself,⁷ as evinced by the wielding of FRTs for bulk identification and verification of protestors across the globe.⁸

With the protection of privacy being largely contingent upon limitations prescribed by state constitutions and the willingness of states to abide by international human rights norms,⁹ it becomes imperative to examine the level of protection accorded to privacy in the domestic sphere. Given the risks of generalising the experience of Western polities, which constitute an unrepresentative sample in an increasingly multipolar world,¹⁰ this chapter seeks to move beyond the United States (US)–European Union (EU) duality and centre its analysis of the implications of state surveillance facilitated through FRTs on the right to privacy in the Global South. Constitutionalism in the Global South, a term with socio-economic and epistemic rather than purely geographical connotations,¹¹ is marked by ‘similar macro-dynamics but also profoundly heterogeneous micro-dynamics’ whilst simultaneously being intertwined with constitutionalism in the Global North, thereby enriching the understanding of the latter as well.¹² As such, an analysis of the right to privacy vis-à-vis FRTs in the Global South entails examining a wide range of regime types, legal cultures, and experiences. To facilitate this analysis, the chapter will first unpack, briefly, the legal protections accorded to privacy and the impact thereof on the prevailing conception of privacy. Thereafter, it will delve into the use of FRTs for the purposes of surveillance and examine the impact thereof on privacy, in particular, and civil liberties, in general. Lastly, it will argue that the emphasis on informational privacy, coupled with uneven levels of protection, allows states to not only infringe upon their citizens’ privacy rights but also clamp down on dissent and democratic participation.

⁷ Monika Zalnieriute, ‘Burning Bridges: The Automated Facial Recognition Technology and Public State Surveillance in the Modern State’ (2021) 22 *Columbia Science and Technology Law Review* 284, 286.

⁸ Office of the United Nations High Commissioner for Human Rights (OHCHR), ‘Report of the UN High Commission for Human Rights on the impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests’ (25 June 2020) UN Doc A/HRC/44/24, 7 <www.ohchr.org/en/documents/thematic-reports/ahrc4424-impact-new-technologies-promotion-and-protection-human-rights>.

⁹ Kristian Humble, ‘International Law, Surveillance and the Protection of Privacy’ (2021) 25(1) *International Journal of Human Rights* 1, 4.

¹⁰ Philip Dann, Michael Riegner, and Maxim Bönnemann, ‘The Southern Turn in Comparative Constitutional Law: An Introduction’ in Philipp Dann, Michael Riegner, and Maxim Bönnemann (eds), *The Global South and Comparative Constitutional Law* (OUP 2020) 1, 4.

¹¹ Florian Hoffman, ‘Facing South: On the Significance of An/Other Modernity in Comparative Constitutional Law’ in Philipp Dann, Michael Riegner, and Maxim Bönnemann (eds), *The Global South and Comparative Constitutional Law* (OUP 2020) 41, 45.

¹² Dann, Riegner, and Bönnemann (n 10) 1, 3.

2 Privacy: Understanding the Prevailing Conception

Privacy has occupied a highly contested position in contemporary politico-legal discourse,¹³ however, the contours of the right are not sufficiently well-defined,¹⁴ with the efforts aimed at defining and conceptualising privacy often described as dissatisfactory.¹⁵ Daniel Slove has identified six conceptions of privacy, namely, '(1) the right to be let alone; (2) limited access to the self; (3) secrecy; (4) control over personal information; (5) personhood; and (6) intimacy'.¹⁶ There is a significant divergence amongst both scholars and legal systems on which of these conceptions is, and ought to be, the prevalent one. The definitional uncertainties that plague the concept of privacy significantly complicate the task of recognising and implementing a universal right to privacy.¹⁷ It has been argued that adopting a comparative lens towards this issue could provide insights on how best to resolve contesting claims by governments and civil society;¹⁸ an objective that becomes pressing not only in the context of the increasing use of FRTs for surveillance and law enforcement but also in light of the global trends pertaining to rising authoritarianism.¹⁹

2.1 The Right to Privacy in Asia: The Emergence of an Informational Conception in a Legally Heterogenous Region

In Asia, it is pertinent to note that the absence of a regional human rights agreement has grounded the human rights discourse in domestic constitutional law.²⁰ As of 2014, only four out of twenty-six countries studied by Graham Greenleaf did not contain any constitutional safeguards pertaining to the right to privacy.²¹ However, several of these constitutions only provide for the protection of certain aspects of the right to privacy, such as the inviolability of the home and correspondence.²² Only a few, such as Nepal and South Korea, explicitly recognise a general right to privacy.²³ In India and Indonesia, despite no explicit textual mention

¹³ Norman Witzleb and others, 'An Overview of Emerging Challenges in Privacy Law' in Norman Witzleb and others (eds), *Emerging Challenges in Privacy Law: Comparative Perspectives* (CUP 2014) 1.

¹⁴ Ronald J Krotoszynski Jr, *Privacy Revisited: A Global Perspective on the Right to Be Left Alone* (OUP 2016).

¹⁵ Daniel J Slove, 'Conceptualizing Privacy' (2002) 90 California Law Review 1087, 1154.

¹⁶ ibid 1087, 1094.

¹⁷ Alexandra Rengel, *Privacy in the 21st Century* (Brill 2013) 28.

¹⁸ Witzleb and others (n 13) 11.

¹⁹ Tom Ginsburg, 'Authoritarian International Law?' (2020) 114(2) American Journal of International Law 221.

²⁰ Michael C Davis, 'The Political Economy and Culture of Human Rights in East Asia' (2011) 1(1) Jindal Journal of International Affairs 48, 49.

²¹ Graham Greenleaf, *Asian Data Privacy Laws: Trade and Human Rights Perspectives* (OUP 2014) 472. These four countries were Brunei, Laos, Sri Lanka, and Singapore.

²² Eg Constitution of Bangladesh (1972), art 43; Constitution of China (1982), arts 39–40.

²³ Constitution of Nepal (2015), art 28; Constitution of the Republic of Korea (1948), art 17.

of 'privacy', the constitutional courts have held the right to privacy to be implicit within existing rights, that is, the right to life or the right to dignity and to feel secure, respectively.²⁴ It is pertinent to note, however, that constitutional guarantees have either not translated into tangible protection of the right to privacy (eg in Bangladesh or Pakistan),²⁵ or the extent of protection is indeterminate (eg in Afghanistan).²⁶ In the context of data privacy, about half of the countries studied by Greenleaf had enacted legislation in relation to the same, of which only six countries had comprehensive data privacy laws covering both the public and private sectors. This is in sharp contrast with the global trend, where 90 per cent of extant data privacy legislations were comprehensive, rather than covering only the public or private sector.²⁷ In the absence of a comprehensive data privacy law, privacy rights can often be protected, albeit to a limited extent, by right to information legislation, which may not only provide individuals with the ability to access and rectify their data but also protect others from accessing the same.²⁸ At least ten jurisdictions in Asia have enacted right to information statutes.²⁹ Despite the privacy landscape often being defined by a patchwork of legislations that do not necessarily deal with privacy in the traditional sense,³⁰ comprehensive data protection legislation has either been enacted recently or is pending in many countries.³¹ This trend assumes significance as the move towards comprehensive data privacy legislation, rather than regulations limited to the private sector, could 'reverse what could otherwise become an Asian rejection of the democratic dimension of data privacy laws represented by coverage of the public sector'.³² The heterogeneity of Asian countries, in terms of both cultures and regime types, represents a challenge in terms of determining a uniform conception of privacy in liberal democratic terms. However, recent trends suggest a consolidation of the law governing the informational aspect of privacy through data privacy legislation rather than overarching interpretive or textual reform at the constitutional level.

²⁴ *Justice KS Puttaswamy (Retd) v Union of India* (2017) 10 SCC 1; Institute for Policy Research and Advocacy and Privacy International, 'The Right to Privacy in Indonesia' (September 2016), 3 <<https://uprdoc.ohchr.org/uprweb/downloadfile.aspx?filename=3914&file=EnglishTranslation>>.

²⁵ Smitha Prasad and Sharngan Aravindakshan, 'Playing Catch Up: Privacy Regimes in South Asia' (2021) 25(1) International Journal of Human Rights 79, 105.

²⁶ *ibid* 79, 101.

²⁷ Greenleaf (n 21) 472.

²⁸ *ibid* 475.

²⁹ Greenleaf (*ibid*) identifies nine jurisdictions, namely: Bangladesh, China, India, Indonesia, Japan, South Korea, Nepal, Taiwan, and Thailand. However, subsequently, Sri Lanka has enacted the Right to Information Act 2016 (Act No 12 of 2016).

³⁰ For instance, in Bangladesh, the right to privacy and state surveillance falls within the ambit of the Telegraph Act 1886; the Bangladesh Telecommunication Act 2001; the Information and Communications Technology Act 2006; and the Digital Security Act 2018. See Prasad and Aravindakshan (n 25) 79, 105.

³¹ Thailand and Sri Lanka have recently enacted the Thai Personal Data Protection Act 2020 and the Personal Data Protection Act 2022 (Act No 9 of 2022), respectively. Data protection bills are pending in India, Pakistan, and Indonesia.

³² Greenleaf (n 21) 478.

2.2 The Right to Privacy in Africa: Data Protection as a Growing Focus

The African Charter on Human and Peoples' Rights (ACHPR) does not explicitly provide for the right to privacy, with the only substantive provision relating to privacy being found in the African Charter on the Rights and Welfare of the Child (ACRWC).³³ The African Cybersecurity and Personal Data Protection Convention has not yet received the requisite support to enter into force, whereas the 2019 African Declaration on Freedom of Expression and Access to Information, which frames the right to privacy in the context of the protection of personal information, is a non-binding interpretive instrument adopted by the African Commission on Human and People's Rights (ACommHPR).³⁴ At the subregional level, the Economic Community of West African States' (ECOWAS) Supplementary Act on Personal Data Protection within ECOWAS constitutes the only binding data protection agreement in Africa, although some other subregional organisations have adopted model laws on data protection.³⁵ At the constitutional level, Greenleaf and Cottier upon an analysis of fifty-five constitutions,³⁶ found that seven constitutions contain specific provisions on the protection of personal data,³⁷ twenty-eight, whilst not addressing data protection specifically, do provide for the right to privacy,³⁸ and an additional twelve provide only for the inviolability and secrecy of correspondence.³⁹ However, the case law on the right to privacy is relatively scant.⁴⁰ In the context of data protection legislation, thirty-two countries have enacted data protection laws, five countries have bills that have not yet been enacted, and seventeen countries have neither laws nor bills in this regard.⁴¹ There appears to be an ascendant trend amongst African countries of enacting domestic data protection legislation, with increasing attempts to harmonise the same at the subregional level. Although the ACHPR does not contain a provision pertaining to the right to privacy, constitutional provisions in numerous states do provide

³³ African Charter on the Rights and Welfare of the Child (adopted 11 July 1990, entered into force 29 November 1999) (ACRWC) CAB/LEG/24.9/49, art 10.

³⁴ Yohannes Eneyew Ayalew, 'Untrodden Paths Towards the Right to Privacy in the Digital Era Under African Human Rights Law' (2022) 12(1) International Data Privacy Law 16, 20–22.

³⁵ Graham Greenleaf and Bertil Cottier, 'International and Regional Commitments in African Data Privacy Laws: A Comparative Analysis' (2022) 44 Computer Law and Security Review 14–17.

³⁶ *ibid* 6.

³⁷ Eg Constitution of Algeria (2020), arts 47–48; Constitution of Cape Verde (1980), arts 38, 41–42; Constitution of Mozambique (2004), arts 41, 71.

³⁸ Eg Constitution of Eritrea (1997), art 18; Constitution of Kenya (2010), art 31; Constitution of Malawi (1994), art 21; Constitution of South Africa (1996), art 14; Constitution of Zimbabwe (2013), art 57.

³⁹ Several of these constitutions provide not only for the inviolability of correspondence but also of the domicile. See eg Constitution of Benin (1990), arts 20–21; Constitution of Ghana (1992), art 18(2); Constitution of Senegal (2001), arts 13, 16.

⁴⁰ Alex B Makulilo, 'The Future of Data Protection in Africa' in Alex B Makulilo (ed), *African Data Privacy Laws* (Springer 2016) 371, 373.

⁴¹ Greenleaf and Cottier (n 35) 4–6.

for the same. It has, in fact, been argued that the constitutional right to privacy is too broad and that a growing body of case law, in countries such as South Africa and Kenya, is increasingly addressing non-informational aspects of the same.⁴² However, a broader conception of privacy, rather than a purely informational one, that sufficiently accounts for autonomy and dignity, could be crucial in providing protections against unlawful state surveillance.

2.3 The Right to Privacy in South America: Constitutional and Statutory Emphasis on Informational Privacy

In South America, constitutions variously provide for the inviolability of the home and the domicile,⁴³ the right to privacy generally,⁴⁴ and the specific right to protection of personal data.⁴⁵ A large number of South American constitutions also provide for the writ of habeas data,⁴⁶ which guarantees individuals the right to access, challenge, and rectify personal information collected by the state or private actors of a public nature.⁴⁷ It has, however, been argued that since the habeas data remedy only provides ex post relief, it offers only a minimal level of protection of an individual's right to privacy.⁴⁸ Over time, several South American countries, have also enacted comprehensive legislation for the protection of personal data.⁴⁹ Furthermore, article 11 of the American Convention on Human Rights (ACHR), provides for a right to privacy,⁵⁰ which has been interpreted by the Inter-American Court of Human Rights (IACtHR) as being a multifaceted right associated with the dignity of the individual that encompasses autonomy.⁵¹ The right to privacy is clearly enshrined at both the constitutional and regional levels in South America,

⁴² Alex B Makulilo, 'Myth and Reality of Harmonisation of Data Privacy Policies in Africa' (2015) 31(1) Computer Law and Security Review 78, 80.

⁴³ Eg Constitution of Argentina (1853), arts 18–19; Constitution of Panama (1972), arts 26, 29; Constitution of Peru (1993), art 2(9)–(10).

⁴⁴ Eg Constitution of Bolivia (2009), art 21(2); Constitution of Nicaragua (1987), art 26; Constitution of Paraguay (1992), art 33.

⁴⁵ Eg Constitution of Colombia (1991), art 15; Constitution of Dominican Republic (2015), art 44; Constitution of Mexico (1917), art 16.

⁴⁶ Eg Constitution of Brazil (1988), art 5(72); Constitution of Ecuador (2008), art 92; Constitution of Peru (1993), arts 2(5)–(6), 200(3); Constitution of Venezuela (1999), art 28.

⁴⁷ Marc Tizoc Gonzalez, '*Habeas Data*: Comparative Constitutional Interventions from Latin America Against Neoliberal States of Insecurity and Surveillance' (2015) 90(2) Chicago-Kent Law Review 641.

⁴⁸ Josiah Wolfson, 'The Expanding Scope of Human Rights in a Technological World—Using the Interamerican Court of Human Rights to Establish a Minimum Data Protection Standard Across Latin America' (2017) 48(3) University of Miami Inter-American Law Review 188, 209–10.

⁴⁹ These countries include Brazil, Chile, Colombia, Costa Rica, Mexico, Nicaragua, Panama, Peru, and Uruguay.

⁵⁰ Organization of American States (OAS), American Convention on Human Rights (ACHR) (22 November 1969) OAS Treaty Series No 36, 1144 UNTS 123, art 11.

⁵¹ *Artavia Murillo et al (In Vitro Fertilization) v Costa Rica* Inter-American Court of Human Rights Series C No 257 (28 November 2012), para 143.

albeit with a degree of emphasis on the informational aspect, particularly at the domestic level, as witnessed through the prevalence of habeas data provisions in state constitutions, as well as an increasing trend of enacting data protection legislation. The wider notion of privacy, as laid down by the ACHR and reflected in constitutional provisions safeguarding a general right to privacy, assumes significance in the context of the use of FRTs, which will be dealt with in section 3.

3 FRTs and the Surveillance State: The Rise of Digital Authoritarianism

Surveillance of citizens by the state is not a new phenomenon.⁵² For instance, historically, human rights violations by South American dictatorships were rooted in the surveillance of dissident elements of civil society by intelligence agencies.⁵³ However, expanding technological capabilities have led to the replacement of labour-intensive surveillance methods, characterised by the surveillance of the few by the many, with methods such as FRTs.⁵⁴ For authoritarian regimes, this entails a reduction of manpower costs as well as the elimination of the risk of security forces utilising the resources empowering them to repress the opposition to act against the regime itself.⁵⁵ For instance, Feldstein argues that AI-fuelled technologies, such as FRTs, form an integral part of the system of control underlying the Communist Party's regime in China.⁵⁶ China has reportedly deployed a network of over 626 million FRT cameras across the country.⁵⁷ Although originally intended for legitimate purposes, there are concerns regarding the increasingly intrusive nature of FRT surveillance,⁵⁸ exemplified by allegations of racial profiling of Uyghur Muslims,⁵⁹ and reports of their subsequent detention in 're-education camps'.⁶⁰ The same is emblematic of the phenomenon known as 'function creep'—the

⁵² Eg Andreas Lichter and others, 'The Long-Term Costs of Government Surveillance: Insights from Stasi Spying in East Germany' (2021) 19(2) *Journal of the European Economic Association* 741, describing how the Ministry for State Security in East Germany 'implemented one of the largest and densest surveillance networks of all time'.

⁵³ Eduardo Bertoni and Collin Kurre, 'Surveillance and Privacy Protection in Latin America: Examples, Principles, and Suggestions' in Fred H Cate and James X Dempsey (eds), *Bulk Collection: Systemic Government Access to Private-Sector Data* (OUP 2017) 325, 326.

⁵⁴ Ian Berle, *Face Recognition Technology: Compulsory Visibility and Its Impact on Privacy and the Confidentiality of Personal Identifiable Images* (Springer 2020) 39.

⁵⁵ Steven Feldstein, 'The Road to Digital Unfreedom: How Artificial Intelligence is Reshaping Repression' (2019) 30(1) *Journal of Democracy* 40, 42.

⁵⁶ *ibid* 40, 48.

⁵⁷ Lindsey Jacques, 'Facial Recognition Technology and Privacy: Race and Gender—How to Ensure the Right to Privacy is Protected' (2021) 23 *San Diego International Law Journal* 111, 135.

⁵⁸ *ibid* 111, 135.

⁵⁹ Paul Mozur, 'One Month, 500,000 Face Scans: How China is using AI to profile a minority' *The New York Times* (14 April 2019) <www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html>.

⁶⁰ 'Who are the Uyghurs and why is China being accused of genocide?' *BBC* (21 June 2021) <www.bbc.com/news/world-asia-china-22278037>.

expansion of a technological system beyond its original, and proper, purpose.⁶¹ Furthermore, Chinese companies are leading exporters of AI-fuelled surveillance technology,⁶² with Gravett opining that, thereby, 'China has also begun to export its model of digital authoritarianism across the globe, including to Africa'.⁶³ For instance, in 2020, Ugandan police arrested over 836 individuals in connection to protests against the arrest of two presidential candidates by utilising a network of CCTV cameras supplied by the Chinese company Huawei, raising concerns that the Ugandan authorities were accessing the system for conducting FRT enhanced surveillance of those opposed to the incumbent regime.⁶⁴ Although the protests in Uganda assumed a violent character,⁶⁵ a report of the United Nations High Commissioner for Human Rights has highlighted the adverse impact of the routine use of FRTs, in the absence of sufficient safeguards, to identify members of peaceful assemblies on the right to privacy as well as the freedom of speech and assembly.⁶⁶ During the 2019–2020 protests in Hong Kong, protestors responded to these risks by wearing face masks, carrying umbrellas, and destroying surveillance towers, leading the government to ban facial coverings.⁶⁷ This points to attempts at opting out of FRT surveillance being construed as indicative of potential criminality, lending weight to the notion that those who seek to avoid being surveilled have something to hide in the context of state surveillance.⁶⁸ The collation of information on peaceful assemblies in the form of mass surveillance, through the use of technologies such as FRTs, is problematic even in cases where the likelihood of abuse is low owing to the chilling effect indiscriminate surveillance may have on political participation.⁶⁹

With respect to the trend of China being a major supplier of surveillance technology, including FRTs, across the globe, it is pertinent to note that this export is not limited to authoritarian states or regimes accused of human rights violations

⁶¹ Bert-Jaap Koops, 'The Concept of Function Creep' (2021) 13(1) *Law, Innovation and Technology* 29, 53. See also the chapter by Kostina Prifti, Alberto Quintavalla, and Jeroen Temperman in this volume.

⁶² Feldstein (n 4) 14.

⁶³ William H Gravett, 'Digital Neocolonialism: The Chinese Surveillance State in Africa' (2022) 30(1) *African Journal of International and Comparative Law* 39, 40.

⁶⁴ Stephen Kafeero, 'Uganda is Using Huawei's Facial Recognition Tech to Crack Down on Dissent After Anti-Government Protests' (*Quartz Africa*, 27 November 2020) <<https://qz.com/africa/1938976/uganda-uses-chinas-huawei-facial-recognition-to-snare-protesters/>>.

⁶⁵ Stephen Kafeero, 'Uganda's Election Run-Up Has Turned Deadly After Opposition Candidates Were Arrested' (*Quartz Africa*, 19 November 2020) <<https://qz.com/africa/1935713/uganda-detains-bobi-wine-kicks-off-deadly-election-violence/>>.

⁶⁶ OHCHR (n 8) 9.

⁶⁷ Tatum Millet, 'A Face in the Crowd: Facial Recognition Technology and the Value of Anonymity' (*Columbia Journal of Transnational law Bulletin*, 18 October 2020) <www.jtl.columbia.edu/bulletin-blog/a-face-in-the-crowd-facial-recognition-technology-and-the-value-of-anonymity>.

⁶⁸ Elia Zuriek, 'Theorizing Surveillance' in David Lyon (ed), *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination* (1st edn, Routledge 2002) 32, 44, where he articulates this idea in the context of surveillance by employers at the workplace.

⁶⁹ Titus Stahl, 'Privacy in Public: A Democratic Defence' (2020) 7(1) *Moral Philosophy and Politics* 74.

such as Zimbabwe and Venezuela but extends to liberal democratic states.⁷⁰ Furthermore, it is not only China but also the US that has emerged as a leading purveyor of surveillance technology powered by AI.⁷¹ Indeed, as mentioned earlier, the deployment of FRTs by governments is becoming increasingly ubiquitous, with their utilisation being prevalent in heterogeneity of countries across all geographical regions and regime types.⁷² For instance, in Africa, FRTs are being used for surveillance purposes in not only the aforementioned cases of Uganda and Zimbabwe but also countries such as Kenya and South Africa, where their usage has raised concerns regarding function creep, for instance, by exacerbating the risk of social sorting,⁷³ which entails the collection of group and personal data by surveillance systems to ‘classify people and populations according to varying criteria, to determine who should be targeted for special treatment, suspicion, eligibility, inclusion, access, and so on’.⁷⁴ The pervasive reach of FRTs is also demonstrated by Al Sur’s report, which maps thirty-eight FRT initiatives across nine South American countries, whilst highlighting the potential of FRTs to entrench structural inequities that have historically impacted the region.⁷⁵ In Asia, the risks posed by the proliferation of FRTs are compounded by the fact that all twenty-three Asian countries studied by Greenleaf have a national identity card system,⁷⁶ which, when combined, *inter alia*, with CCTV footage, can potentially reduce citizens to data subjects and facilitate panoptical surveillance.⁷⁷ The anxieties surrounding the usage of FRTs for problematic purposes in Asia came to the fore in the case of Myanmar, where protestors voiced fears of being subjected to surveillance through FRTs by the Military Junta.⁷⁸

⁷⁰ Feldstein (n 4) 14.

⁷¹ *ibid* 8.

⁷² *ibid* 7–8.

⁷³ Karen Allen and Isel van Zyl, ‘Who’s Watching Who? Biometric Surveillance in Kenya and South Africa’ (*Enact*, 11 November 2020) <<https://enact-africa.s3.amazonaws.com/site/uploads/2020-11-11-biometrics-research-paper.pdf>>.

⁷⁴ David Lyon, ‘Surveillance as Social Sorting: Computer codes and mobile bodies’ in David Lyon (ed), *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination* (1st edn, Routledge 2002) 13, 20.

⁷⁵ Al Sur, ‘Facial Recognition in Latin America: Trends in the Implementation of a Perverse Technology’ 7 <https://www.alsur.lat/sites/default/files/2021-10/ALSUR_Reconocimiento%20facial%20en%20Latam_EN_Final.pdf>.

⁷⁶ Greenleaf (n 21) 475.

⁷⁷ Berle (n 54) 159; Michael Foucault, *Discipline and Punish: The Birth of the Prison* (2nd edn, Vintage Books 1995) 202, outlines how the Panopticon was a machine envisaged by Jeremy Bentham that allowed the automation and dis-individualisation of power by ‘dissociating the see/being seen dyad: in the peripheric ring, one is totally seen, without ever seeing; in the central tower, one sees everything without ever being seen’.

⁷⁸ Rina Chandran, ‘Fears of “Digital Dictatorship” as Myanmar Deploys AI’ (*Reuters*, 18 March 2021) <www.reuters.com/world/china/fears-digital-dictatorship-myanmar-deploys-ai-2021-03-18/>.

4 The Need for a Broader Conception of Privacy: Moving Beyond the Purely Informational

In authoritarian regimes, surveillance technologies such as FRTs have led to the development of a model of digital authoritarianism. Although democracies ostensibly utilise FRTs for legitimate purposes such as the prevention of terrorism and heinous crimes,⁷⁹ the use of FRTs presents a ‘high risk, high reward scenario’ that entails the risk of moving towards a police state despite its expansiveness and cost-effectiveness.⁸⁰ With existing legislation regulating surveillance providing insufficient safeguards in relation to FRTs,⁸¹ the right to privacy, as enshrined in state constitutions and regional human rights instruments, could play a crucial role in mitigating the function creep plaguing FRTs. The absence of such checks and balances can facilitate the flourishing of digital authoritarianism, as evinced by the case of China, especially considering that there is no regional human rights agreement in Asia. Furthermore, China has neither ratified the International Covenant on Civil and Political Rights (ICCPR), article 17 of which provides for the right to privacy, nor are its constitutional provisions regarded as justiciable.⁸² This, in turn, has provided impetus to the growth of a surveillance state, fuelled by technologies such as FRTs. Likewise, even where independent constitutional courts enforce justiciable constitutional provisions, protecting only a limited aspect of privacy, such as the inviolability of the home or correspondence, or a greater emphasis on data protection, as in the case of habeas data provisions in South America, can result in insufficient protection against the intrusive use of FRTs.

The heavy emphasis on data protection, which can be seen at the constitutional level in South America, may also be witnessed vide the mushrooming of data protection legislation in South America itself, Africa, and Asia. Although privacy is a multidimensional concept, it is its informational aspect that has arguably emerged as ‘the most significant and contested contemporary form’⁸³

Data protection legislation tends to safeguard the informational aspect of privacy, which not only conflates privacy with confidentiality or secrecy but applies to a reserved private sphere that ought to be free from intrusion.⁸⁴ FRTs, on the other hand, impose compulsory visibility upon individuals subjected to their

⁷⁹ Königs (n 6) 8–9.

⁸⁰ Ameen Jauhar, ‘Facing up to the Risks of Automated Facial-Recognition Technologies in Indian Law Enforcement’ (2020) 16 *Indian Journal of Law and Technology* 1, 4.

⁸¹ Grace Mutung’u, ‘Surveillance Law in Africa: A Review of Six Countries (Kenya Country Report)’ (2021) *Institute of Development Studies* 91; Grace Mutung’u, ‘Surveillance Law in Africa: A Review of Six Countries (South Africa Country Report)’ (2021) *Institute of Development Studies* 172.

⁸² Greenleaf (n 21) 196–97.

⁸³ Gary Marx, ‘Coming to Terms: The Kaleidoscope of Privacy and Surveillance’ in Beate Roessler and Dorota Mokrosinska (eds), *Social Dimensions of Privacy: Interdisciplinary Perspectives* (CUP 2015) 36.

⁸⁴ Berle (n 54) 79; Manuel José Cepeda Espinosa, ‘Privacy’ in Michel Rosenfeld and Andras Sajó (eds), *Oxford Handbook of Comparative Constitutional Law* (OUP 2012) 966, 967.

gaze,⁸⁵ which despite creating visibility, prevents individuals, or members of targeted groups, from controlling their public identity.⁸⁶ This highlights the decisional aspect of privacy, which is intrinsically linked to autonomy.⁸⁷ The fact that this violation takes place in the public sphere, coupled with the role of privacy as a ‘precondition for … democratic, participatory friction’,⁸⁸ highlights the shortcomings of a purely informational conception of privacy. Data protection laws, grounded in the notion of privacy as informational control, have been unable to prevent the rollout of disruptive technologies, such as FRTs, that may facilitate surveillance that has a debilitating effect on the sociopolitical realm.⁸⁹ More significantly, a purely informational conception of privacy prevents the entrenchment of a broader formulation that safeguards human dignity and democratic freedoms against surveillance by the state, and shirks foundational questions on whether surveillance ought to be tolerated by focusing only on the limits and control of surveillance, if it is allowed.⁹⁰ Safeguarding privacy in the public sphere, particularly public behaviour, assumes significance in light of the value of such behaviour towards bolstering ‘collective autonomy and democratic self-government’,⁹¹ especially since the legitimacy of the political process is contingent upon deliberation within the public sphere.⁹² However, the decisional autonomy required for engaging in meaningful deliberation within the public sphere may significantly be stymied by virtue of mass surveillance. To that end, an overarching general right to privacy, which protects not only the informational, and local, but also the decisional aspects of privacy,⁹³ could allow courts to adjudicate the right in a manner that can check instances of potential function creep that hinder meaningful political participation. At present, Asia does not have a binding regional human rights agreement whereas the ACHPR does not contain a substantive provision pertaining to privacy, the broad conception of privacy under article 11 of the ACHR could be wielded to safeguard the right to privacy in a manner that can insulate political participation

⁸⁵ Berle (n 54) 79.

⁸⁶ Scott Skinner-Thompson, ‘Agonistic Privacy & Equitable Democracy’ (2021) 131 *Yale Law Journal Forum* 459.

⁸⁷ Marjolein Lanzing, ‘Strongly Recommended: Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies’ (2019) 32 *Philosophy & Technology* 549, 558.

⁸⁸ Skinner-Thompson (n 86) 457.

⁸⁹ Valerie Steeves, ‘Reclaiming the Social Value of Privacy’ in Ian Kerr, Valerie Steeves, and Carole Lucock (eds), *Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society* (OUP 2009) 191–92.

⁹⁰ *ibid* 192–95.

⁹¹ Stahl (n 69) 75.

⁹² *ibid* 82–84.

⁹³ Beate Roessler, ‘Three Dimensions of Privacy’ in Bart van der Sloot and Aviva de Groot (eds), *The Handbook of Privacy Studies: An Interdisciplinary Introduction* (AUP 2018) 137, where she identifies and differentiates three dimensions of privacy: (i) decisional privacy, which aims to secure autonomy of the individual not ‘just in the intimate sphere, but in private acts and behaviour in public too’; (ii) informational privacy, which includes protection of individual freedom, in this case information of the person, even when it involves something that happens in public; and (iii) local privacy, which is similar to the traditional notion of privacy, that is, privacy in one’s own home.

from the intrusive use of FRTs for state surveillance. Meanwhile, throughout the Global South, embedding this right at the constitutional level as an integral guarantee prevents not only the executive, that is, administrative and police authorities, but also the legislature from violating the same,⁹⁴ thereby preventing the state from legalising the intrusive usage of FRTs through surveillance statutes.

5 Conclusion

FRTs are being utilised for surveillance purposes by a heterogeneity of countries across geographical regions and regime types in the Global South. The deployment of FRTs for surveilling the population, engaging in social sorting, and curbing dissent, has resulted in the face, which has historically been wielded by civil society for the purposes of political mobilisation, emerging as a political tool against democratic participation. The pervasive use of FRTs could, thus, accentuate trends of democratic backsliding. Their deployment in public spaces, coupled with potential effects on meaningful political participation, highlights the need for moving beyond a purely informational conception of privacy, which is not only safeguarded at the constitutional level in much of South America but is also becoming increasingly entrenched throughout the Global South through the proliferation of data protection legislation. Consequently, a right to privacy that also protects the decisional aspect of privacy, which, in turn, underscores individual autonomy, must be guaranteed at the constitutional level. This may entail either meaningfully interpreting existing provisions and human rights instruments providing for a general right to privacy or effecting constitutional reforms, where privacy is insufficiently protected in the text of the constitution. Given that it is the governments that are increasingly deploying FRTs and other AI-fuelled surveillance technologies, it may fall upon constitutional courts to innovatively interpret and safeguard the right to privacy in a broader sense, so as to prevent political participation from being stifled due to the disruptive effects of such technologies.

⁹⁴ Espinosa (n 84) 966–67.

The Production of and Control over Data in the AI-Era

The Two Failing Approaches to Privacy Protection

Bart van der Sloot

1 Knowledge Production in the Twenty-First Century

Suppose it is 1970. A person or organisation wants to obtain knowledge on person A, to take an informed decision. How does the person or organisation in question produce knowledge? There are a few things that can be presumed, as they are logically evident. Person A can, for example, logically speaking not be at place 1 and place 2 at the same time. Person A cannot be a married bachelor. Person A cannot be person B. Then there is empirical knowledge. Person A goes to many jazz concerts. Person B says that person A is unreliable. People around person A all are supporters of a left-wing political party. On the face of it, person A is around the age of thirty. Sixty-five per cent of people between fifteen and thirty years of age think the band, Pink Floyd, is awesome.

Obviously, neither of these knowledge sources is absolutely reliable. Even the logical presumption may be challenged or altered over time. Though it was considered logically contradictory that a person be both a man and a woman, now that is no longer an absolute.

From an empirical data perspective, it is clear that what people say about themselves is often unreliable, both because they are purposefully selective, because they cannot possibly tell everything in all its nuances and because human memory is fallible. Similarly, what people say about others is imprecise and misguided. Observations may be more reliable, but there are limits to what one can observe; also, there are choices in how to categorise the data and the fact that a person is or knows to be observed may influence her behaviour, while another problem is that observation data does not give an indication as to the meaning of the data. What does the fact that person A often goes to a jazz concert hall mean? Presumably, she likes jazz, but maybe she does not and just has a friend whom she joins, or maybe she works at the jazz hall.

Now suppose person A has moved from the countryside to a big city far away and her new friends want to find out about her past, present, and future. How do

they arrive at knowledge? How do the people she meets know who she is and what she is like? They have to go on what she says, primarily, and what she does. This allows person A a large room to reinvent herself. She may ignore part of her past, emphasise specific aspects of her present personality and start acting differently than she did in the city she grew up in. She might wear a different outfit, put on a different accent, or start visiting rock concerts, which she never did before. One might say that the way she presents herself is false, though person A herself may feel she is foreshadowing her new self, or perhaps even her 'real' self. How convincing her new self is, depends primarily on herself. Can she hide her thick countryside accent? Does she know how to wear the clothes that come with her new fashion style? Does she know enough about rock music to keep up conversations with her new friends? Other sources are difficult to find. There are very few sources that are available on her. Her friends and family are in another part of the country, and though it is possible to give them a call or visit them, there is a practical barrier to doing so. Objects providing cues are fewer still, though there might be a diary, a school photo, or some concert tickets in the basement of the parents of person A.

Now, suppose Person A has joined the soccer team in her new city. In a year, she has grown to be one of the most reliable goalkeepers in the team. Person B has to take a penalty with A performing the role of goalkeeper. How can person B know which corner to choose? Perhaps the most effective option for person B is to choose randomly, given that neither B knows what A will do nor A what B will do, or, because A most likely knows nothing about B, choose her own favourite corner. B could try to find out about A's presumed actions. The sources are varied this time. She could ask A, of course, but slim chances that A will tell, though perhaps she might do so in a bar, where A does not know that she will be confronted with B on the field at some point in time. More telling probably will be watching the behaviour of A. But because there are no recordings available, this would require B to visit quite a number of training sessions and matches, more so because the number of penalties taken is low. Also, if person A sees person B watching her, she may start deceiving her by acting differently. Perhaps, her best shot of obtaining knowledge is asking a former teammate of A's. But, even on the off chance that she finds someone who wants to disclose information about A, for example, because she really likes B or because she is angry with A, the chances that C has closely and correctly observed A's preferences for corners are slim. In addition, there is always the chance that C deceives B. Alternatively, B could study other goalkeepers and find that most will dive for the left upper corner and very few will stand still to catch a ball that is shot straight at the goalkeeper.

Finally, suppose person A is applying for a job and company D wants to assess how person A will develop over time before deciding whether to give her the job. What does it have to go on? In this scenario, most information sources are available. They can go on what a person says—I want to become a leader in my field; I am happy to follow orders—and how she acts during the interview. There may be

informational cues from others, for instance, a former employee who can tell how person A has performed. There may be objective sources, such as a university degree. And the employer can extrapolate from the behaviour of others: 'Of all people in the group to which person A belongs, we have experienced that 90 per cent does not perform well.' Neither source may be fully reliable, but given the multitude of sources, it becomes possible to crosscheck them, though it has to be stressed that the production of knowledge is influenced by the way data are gathered, by whom, and how they are structured. For example, when A is a young woman and the job interview is held by four white men nearing sixty years of age, not only will they interpret data differently, A will also likely behave differently than when the job interview was performed by a different group of people.

As a side note, but an important one nonetheless, it should be stressed that the main system for representing knowledge will be natural language. Natural language has implicit presumptions and inherent ways of categorising information. Both because of the use of natural language and the limits of storing and categorising data in the 1970s, there was a tendency to rely primarily on qualitative data and only marginally on quantitative data, except for a small number of circumstances where databases were used.

Suppose it is 2020 now. The same situations apply. There is a huge shift in the knowledge available for each of the scenarios and in the way they are presented.

For the first scenario, there will be more information available both about the environment where a person grew up and what the person did in the past, how she dressed, what type of activities she engaged with, and so on. Photos, videos, blogs, and tweets are often available online and difficult to remove from the web when a person so desires, if only because many of her photos will be on webpages of her friends or in the hands of third parties, such as concert halls. This allows A's new friends to verify more easily whether the story she sells is coherent and consistent.

An additional shift is that such data is mediated by tech companies, such as Apple, Facebook, and Google. They have the technical facilities to store data and distil patterns out of them. A person herself is limited in terms of harvesting patterns, although some people keep detailed diaries, very few people minutely write down their whereabouts, their activities, and feelings; even the quantified self-movement is mediated by tech firms. These tech companies can distil detailed patterns: the number of times a person went to McDonalds' after school and what impact her period and the weather had on that pattern, her presumed emotional state derived from activities and music played per minute, among many other things. These tech firms will not change their profile of A merely because she has moved to a new city. What is important is the rise of inferred data. From two data points, another data point may be inferred, for example, 'when a person listens to a certain category of music, she is likely to be gay'.

When assessing which corner A is likely to choose as a keeper, there will be a wealth of information available. There may be pattern data on person A's

preferences: again, her preferences may be calculated and diversified on the basis of other data points, such as the weather, the moment in the game that the penalty is taken, the colour of the shirt of the penalty taker, and so forth. In addition, there may be data on A's peers and detailed information on A's physiology and psychology may be obtained. Cameras can be used in real time to check, for example, a person's heart rate, pupil dilation, and micro-expressions not visible to a human being.

A few important changes arise here vis-à-vis the 1970s. First, when being observed in the past, usually this was known to a person, certainly if it concerned longitudinal observation (where multiple observations of the same subjects over a period of time are made). Second, because of the costs and efforts, a selection was made on the person on whom to gather data and the type of data needed. Given the costs and availability of technology, such is no longer the case. Third, observational technologies are omnipresent and ambient. Hence, it is no longer so that data are gathered on selected, pre-identified individuals. Rather, these technologies observe everyone, for example, everyone that walks in a street, enters a certain building, or visits a certain website. Finally, the room for manipulating behaviour has radically changed. Nudging technologies are applied in private and in public domains, by governmental agencies and companies, making use of the detailed psychological insights gained through the extensive data gathering.

Finally, when selecting a candidate for a job, generally, there will be abundant data on person A and data about A's peers. Two important shifts take place here. First, increasingly many parts of the job selection process are automated, such as a preselection of application letters, the choice as to whom to invite for a job interview and even the eventual selection is increasingly done fully automated, with the candidate having to fill out forms and perform tests, that lead to a certain performance rate and psychological profile. Automation in this sense replaces human biases with the selection committee for data biases and algorithmic bias. Second, predictive profiles are used not only to make the best decision, but also to make decisions as to which tasks a person will get, the training she should receive, and so on. Though data processing is often described as enabling personalised decision-making, in fact, what we are dealing with tends to be group profiling; consider for example the following claim, '90 per cent of the persons to which person A belongs had trouble with exerting leadership, so A should receive leadership training'. Thus, the data on her peers is used to make determinative decisions on A's course of life.

Though scenarios are anecdotal in nature, they do show a number of shifts have occurred in the way knowledge is produced in the course of fifty years. Examples include, but are not limited to:

- natural language > formal language;
- qualitative data > quantitative data;
- incidental data > pattern data;

- individual data > group/category data;
- self-narrative > observed data;
- actual data > inferred data;
- assumptions > probabilities.

Having briefly charted the basic shifts in knowledge production in the last few decades, this chapter will now proceed to outline two basic strategies for privacy and data protection as engrained in law (section 2). It will discuss the two ideal models that have inspired the American and the European regulatory regime. By way of example, this chapter will use the European Union (EU) General Data Protection Regulation (GDPR) as reference point, because it is the most advanced data protection regime around the world. However, these discussions are illustrations of issues that arise in other regional and universal data protection regimes as well. For example, article 5 of the GDPR lists the fair information principles that form the backbone of data protection regulation around the world and is also engrained in, for example, the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. Section 3 will show that these two strategies are difficult to uphold vis-à-vis the new ways of knowledge production section 3. Finally, Section 4 will argue that, in order to be viable in the twenty-first century, the privacy regime has to be fundamentally reconceptualised section 4.

2 Two Strategies of Privacy and Data Protection

There exist two ideal-typical models of data regulation. The first one is to give control to individuals over their personal data. In its most extreme form, scholars have suggested that if individuals would gain property or other control rights over their data,¹ they would be able to adequately represent and protect their own interests against the multinationals and governmental organisations that intend to use their data.² This model has clear advantages as it grants citizens autonomy over their personal data and steers away from any form of paternalism. In addition, it sets no absolute boundaries on what organisations can and cannot do with personal information but connects that question to each individual's preferences, which may vary significantly per person.³

But this model also has clear disadvantages. The capacity of citizens to make choices according to their best interests is limited in practice both because of the

¹ Pamela Samuelson, 'Privacy as Intellectual Property?' (2000) 52 Stanford Law Review 1125.

² Min Mun and others, 'Personal Data Vaults: A Locus of Control for Personal Data Streams' (*Proceedings of the 2010 ACM Conference on Emerging Networking Experiments and Technology*, Philadelphia, November–December 2010).

³ Christophe Lazaro and Daniel Le Métayer, 'Control Over Personal Data: True Remedy or Fairy Tale?' (2015) 12 SCRIPTed 3.

complexity of most contemporary data-driven processes involving biometric data, artificial intelligence (AI), and profiling because of the multitude of processes which contain the data of an average citizen, and because of the information-asymmetry between data-driven organisations and the average citizen.⁴ In addition, many of the data-driven processes affect large groups in society or the population in general; leaving it to each and every individual citizen to assess such processes and their potential flaws individually would mean a privatisation of structural problems and would result in well-educated citizens protecting their personal data better than would already marginalised groups.⁵

A second model is to rely on legal standards and governmental enforcement of those standards. Just like there are minimum safety requirements for cars, there are minimum requirements for legitimately processing personal data. It is not left to citizens to assess whether these rules are met, but to an independent governmental organisation, which has the authority to both investigate data-driven organisations and set sanctions and fines when data controllers violate the rules. This means that legal protection is provided to citizens, without them having to assess the validity, legality, and desirability of each individual data process that contains personal data.

However, this model too has its particular disadvantages. Companies stress that legal standards are oftentimes too restrictive, citizens may be limited in having their data processed against their will and legal standards are often too general, absolute, and inflexible and easily become outdated in the constantly developing data-driven environment.⁶ In addition, it is practically impossible for one governmental organisation to assess all data processing operations⁷ and difficult to ensure that parties based in other territories adhere to national standards. This means that supervisory organisations, such as Privacy Commissioners and the Data Protection Authorities (DPAs) in Europe, usually only focus on the bigger data processing operations, that have the biggest potential impact.

Privacy as fundamental right is recognised in most international and regional human rights frameworks, such as—but not limited to—the Universal Declaration on Human Rights (UDHR), the International Covenant on Civil and Political Rights (ICCPR), the European Convention on Human Rights (ECHR), and the American Convention on Human Rights (ACHR). Some regional human rights frameworks, however, do not. The EU Charter of Fundamental Rights is unique in the world in that it recognises both privacy and data protection as independent fundamental rights. It specifies under article 7 ‘Everyone has the right to respect for his or her private and family life, home and communications,’ while article 8 holds:

⁴ Fred H Cate and Viktor Mayer-Schönberger, ‘Notice and Consent in a World of Big Data’ (2013) 3 *International Data Privacy Law* 67.

⁵ Marjolein Lanzing, ‘The Transparent Self’ (2016) 18 *Ethics and Information Technology* 9.

⁶ Tal Z Zarsky, ‘Incompatible: The GDPR in the Age of Big Data’ (2016) 47 *Seton Hall Law Review* 995.

⁷ Colin J Bennett, *Regulating Privacy* (Cornell UP 2018).

1. Everyone has the right to the protection of personal data concerning him or her.
2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.
3. Compliance with these rules shall be subject to control by an independent authority.

There is a discussion on the right interpretation of both of these provisions, which boils down to the question: should they primarily be seen as laying down obligations on the person or organisation that want to access the private life of a citizen or process her data, or should it primarily be seen as a right to control of the citizen who to allow access to her private life or personal data?

With respect to the right to data protection, reference is often made to article 5 of the GDPR, the main law regulating the processing of personal data in the EU. It holds that personal data should be processed lawfully, fairly and in a transparent manner; collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes; adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed; accurate and, where necessary, kept up to date; kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed and processed in a manner that ensures appropriate security of the personal data. These are all obligations posed on the data controller, the party processing personal data on citizens, that are applicable independent of any rights being invoked by them.

On the other hand, many rights by data subjects in the data protection regimes are emerging. The GDPR includes the right to access, the right to copy, the right to information, the right to object, the right to erasure, the right to rectification, the right to data portability, the right to restrict, the right not to be subject to automated decision-making (ADM) and the right to file a complaint. Many data protection or, what is called, informational privacy regimes around the world are focused on what is called informed consent models, where the individual is perceived as being in control of her data and making rational choices over who to grant access to the data for which purposes on the basis of relevant information and in Europe, especially due to German influence, the notion of informational self-determination has become increasingly popular in Europe.⁸ Thus, some argue that, rather than obligations posed on data controllers, the rights of data subjects are the core of the data protection regime.

⁸ Gerrit Hornung and Christoph Schnabel, 'Data Protection in Germany I: The Population Census Decision and the Right to Informational Self-Determination' (2009) 25 Computer Law & Security Review 84.

The same discussion plays a role within the human rights framework in general and the right to privacy in particular. The ECHR, in article 8, provides protection to the right to privacy. Initially, citizens were not allowed to submit a complaint to the European Court of Human Rights (ECtHR) themselves.⁹ Member states could submit inter-state complaints, or a member state or the European Commission on Human Rights (ECommHR) could send an ‘individual submission’ to the Court when they were convinced that that claim had a broader significance, transcending the particularities of that specific matter.¹⁰ It was believed that the majority of the cases would be inter-state complaints and that the individual cases would be brought by legal persons, for example, civil society organisations, and groups (the ECHR was adopted in the wake of the Second World War, where abuse policies were directed at groups and individuals solely treated as part of those groups). Inter-state complaints do not, in fact, revolve around harm claimed by an applicant but concern a general policy or legal system that is deemed to be in violation of the ECHR.¹¹

The ECHR was intended to address larger societal concerns over abuse of governmental power by totalitarian regimes: not the individual harmed through a specific action by the executive branch, for example, a governmental official unlawfully entering a home or wiretapping a telephone, but rather the stigmatisation of minorities, Stasi-like governmental surveillance, and other widespread or systematic anti-democratic practices.¹² Hence, emphasis was placed on negative obligations for states, that is not to abuse their powers, and negative rights for citizens, that is not to be interfered in their rights, rather than on subjective claim rights for natural persons to protect their personal interests in concrete cases and on positive obligations for states to help citizens to pursue their desired life path.¹³

Over time, the ECHR structure and rationale has been changed. Natural persons have been allowed direct access to the ECtHR, the possibility of inter-state complaints never gained traction¹⁴ and the ECtHR has barred groups from submitting a claim as a group¹⁵ and has been hesitant to allow legal persons to invoke the right to privacy, as it feels that this doctrine primarily provides protection to

⁹ Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR), art 48.

¹⁰ Bart van der Sloot, *Privacy as Virtue* (Intersentia 2017).

¹¹ Arthur Henry Robertson, *Collected Edition of the ‘Travaux Préparatoires’ of the European Convention on Human Rights = Recueil des Travaux Préparatoires de la Convention Européenne des Droits de l’Homme* (Nijhoff 1975) 270.

¹² Pieter van Dijk and others, *Theory and Practice of the European Convention on Human Rights* (Intersentia 2006).

¹³ Carolyn Evans, *Freedom of Religion under the European Convention on Human Rights* (OUP 2001).

¹⁴ See Protocol No 9 to the ECHR (1990). Protocol 9 has been repealed as from the date of entry into force of Protocol No 11 to the ECHR (1998), restructuring the control machinery established thereby. Since its entry into force on 1 November 1998, this Protocol forms an integral part of the ECHR, as amended by Protocols Nos 11 and 14.

¹⁵ It does allow citizens that all have been harmed by a specific governmental practice to bundle their complaints.

individual interests and not to societal ones.¹⁶ In addition, the ECtHR has adopted a very tight approach to assessing individual claims: a natural person needs to substantiate concrete, substantial, and individualisable harm that has already materialised and bears a causal relation to the matter complained of.¹⁷ Furthermore, the ECtHR has chosen to take a case-by-case approach, therewith choosing to provide a solution in the concrete circumstances of the case, rather than focusing on general legal questions that have significance for other cases or society as a whole.

Hence, yet again, two interpretations of the right to privacy come to the fore: there are those that primarily see the right to privacy as a subjective claim right attributed to natural persons to protect their private interests and there are those that see human rights, including the right to privacy, as primarily putting an obligation on states not to abuse their power and setting out limits and conditions for the use of power.

3 The Failure of the Two Strategies in the Twenty-First Century

Both strategies work only marginally well in the age of AI with new means of knowledge production.

The above regulatory approach, that of laying obligations on data controllers and having governmental agencies enforce those rules encounters a number of problems. First, even though DPAs in the EU are among the best equipped in the world, they do not have the power or the resources comparable to many of the bigger tech firms. In addition, they lack the technical expertise that these companies have, if only because these Authorities have difficulty to find technical experts that want to work for them because of the lower wages and the less creative aspect of the work compared to the expert work for the companies in question.

Another development that ties into this is that tech companies and spy shops make available to citizens many instruments to collect data about themselves and others. The smartphone, drones, smart doorbells, security cameras, pens with built-in microphones, infrared sensors, and software that is explicitly marketed for surveilling a spouse or employee are all available to consumers and can be ordered online. This has made every citizen into a potential ‘Little Brother’.¹⁸ The democratisation of data technologies has meant that the number of actors gathering data has increased exponentially over the years and the amount of data made available

¹⁶ *Church of Scientology of Paris v France* App no 19509/92 (ECommHR, 9 January 1995).

¹⁷ See eg *Lawlor v United Kingdom* App no 12763/87 (ECtHR 14 July 1988); *Tauria and Others v France* App no 28204/95 (ECommHR, 4 December 1995); *Asselbourg and 78 Others and Greenpeace Association-Luxembourg v Luxembourg* App no 29121/95 (ECtHR, 29 June 1999).

¹⁸ José F Anderson, ‘Big Brother or Little Brother-Surrendering Seizure Privacy for the Benefits of Communication Technology’ (2012) 81 Mississippi Law Journal 895.

on the world wide web grows by the day. This means that no governmental organisation is able to assess all the thousands or millions of photos, videos, blogs, and other material that is put on the web every single day and assess this material on their legitimacy. In addition, they do not have the capacity to pursue every minor privacy violation (eg a person not having been informed by her friend that a photo of her would be published on Instagram) which will lead and has already led to a normalisation of ‘everyday’ privacy violations.

It is questionable whether the solution to grant more power and resources to DPAs is most desirable, because that would mean that the government would essentially closely monitor the activities of citizens, companies, and governmental agencies alike. In fact, this may well lead to a totalitarian regime. The alternative of trusting data controllers to assume their moral (as well as legal) responsibilities is unlikely to work, both because there is little historical evidence to believe they would and because an organisation that is willing to bend the moral and legal boundaries will have a competitive advantage.

An increasing problem for DPAs is that privacy and data protection regimes focus on providing protection to private interests of natural persons and take as anchor ‘personal data’, which means any information relating to an identified or identifiable natural person.¹⁹ Many of the modern processing techniques, however, do not revolve around individual or personal data but are based on aggregated and statistical data on groups and categories. These data processing practices, in principle, fall outside of the competence of the authorities, while decisions made on the basis of aggregated data can have a significant impact on groups and society as a whole.

Finally, a limitation that the first regulatory model cannot account for is that each person has individual privacy preferences. It is true that the world is increasingly datafied and that peoples’ whereabouts, feelings, and preferences are available to tech companies that can make use of the data, but not everyone seems to mind this. Some people see advantages in being datafied, either on an individual level, for instance, the smart refrigerator ordering precisely the type of soya milk that a person likes or the smart TV selecting the movie that perfectly fits a person’s preference; or on societal level consider, for example, people that want to share data about their medical conditions in the hope that this may help the treatment of others. Without individual consent playing a role in the regulatory regime, individual preferences cannot be taken into account, leading to, some would argue, a form of data paternalism.

The alternative regulatory model also reaches its end in the twenty-first century. Consent under the GDPR is defined as follows: ‘consent of the data subject

¹⁹ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (GDPR), art 4.

means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.²⁰ A separate article provides:

Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data. If the data subject's consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. Any part of such a declaration which constitutes an infringement of this Regulation shall not be binding. The data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. Prior to giving consent, the data subject shall be informed thereof. It shall be as easy to withdraw as to give consent. When assessing whether consent is freely given, utmost account shall be taken of whether, inter alia, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract.²¹

The question is to what extent it is realistic that any of these requirements are fulfilled in the context of AI. Artificial intelligence systems are highly complex systems that are trained on historical data, often to make decisions or predictions about the future. They can be deployed in various areas, such as the medical sector (diagnoses and 'personalised' medicine), the law enforcement sector (predictive policing), maintaining public order (eg smart cities that deploy behavioural nudges to make citizens in night life areas less aggressive), in job markets (eg to make a preselection of job applicants), and in warfare (eg to deceive an opponent through the use of deepfakes). Artificial intelligence systems are complex and generally self-learning so that even programmers do not always understand the how and why of the outcomes of AI systems. These systems often use a high number of data and make decisions/predictions on the basis of a complex scoring and weighing of these data and the properties that are assigned to them. That is why the requirements for consent may be difficult to uphold:

- *Informed*: There will always be an information asymmetry between the party asking and the party giving consent. Citizens cannot be asked to understand

²⁰ ibid.

²¹ GDPR, art 7.

what is done with their data in the increasingly complex information systems. The shift from natural to formal language plays into this.

- *Free*: To what extent citizens are truly giving free consent can be questioned. Most services on the internet are either explicitly or practically unavoidable. In addition, structural and ambient monitoring takes place in smart cities, which citizens cannot be asked to avoid.
- *Specific*: In order to be valid, consent should be specific, but this is often not the case. What makes new information technologies so valuable is that it is possible to reuse data for new purposes and that new information can be inferred from existing data. In addition, technological capabilities change at rapid pace.
- *Withdraw*: The right to withdraw only applies to the specific data that were subject to the consent. Inferred data, aggregated data, and longitudinal profiles made by organisations do not fall under the scope of the right to withdraw.

Besides these specific questions as to consent, there are a number of further questions with respect to giving control to citizens and the extent to which that works in the AI environment:

- *Power imbalance*: To give people rights and the power to control their data presumes that they can go to court. Yet, citizens often do not stand a chance against the governmental organisations and large companies that process their data, given the imbalance in knowledge, manpower, time, and resources available.
- *Practical limits*: Even if this power imbalance would be overcome, there are still practical limits for the citizen. It is impossible for anyone to carefully assess the dealings of the more than 5,000 organisations that have data on the average citizen.
- *Ambient and opaque*: Data collection is increasingly ambient, part of the environment, so that people do not notice it; other forms of data collection are covert. People who are not aware that they are affected by data collection practices will not invoke their right. In addition, when going to court, claimants generally have to demonstrate individualisable harm, while modern-day data collection systems revolve around groups and large parts of the population.
- *Nudging*: An additional problem is that what the tech companies invest in is understanding and uncovering psychological and subconscious processes. These processes are highly influential for decision-making processes by human beings but are largely unclear to themselves. This means that the ways for these companies to influence and steer persons' behaviour are increasingly vast. If these organisations are successful, they may even influence our desire to exert rights.

This means that, for a long time, two strategies to privacy and data protection have been deployed: control over personal data by citizens or legal obligations on data controllers with governmental control. Both worked relatively well, often in a mixed form, for decades. Yet, new data technologies that have emerged since the beginning of this millennium and the data infrastructure that exists today make both strategies virtually unable to provide adequate protection in the twenty-first century. Citizens do not have the knowledge, power, and time to control their data and an additional problem is that their autonomy is affected by modern data practices. Governments do not have the knowledge, power, and time to control all data processes and an additional problem is that if they would, this might lead to a totalitarian infrastructure. That is why a reconceptualisation of the right to privacy and privacy protection may be necessary.

4 Conclusion

This chapter has shown that two models of data protection have been developed in the twentieth century. Though they have worked well over the past decades, they face increasingly substantial problems when applied to the way knowledge is produced and controlled in the twenty-first century. Though most regulatory regimes offer a combination of the two regulatory alternative regimes presented above, this has not meant that each approach cancels out the negative aspects of the other. In addition to the challenges discussed in section 3, a final problem should be underlined, namely the fact that tech companies increasingly invest in gathering data about a person's subconsciousness, keep more and more detailed data about their past, more than a person could even do themselves, and are able to make increasingly precise projections of their future life, which are certainly more accurate than the predictions a person can make about their own life.

This means three things:

First, the right to control does not work with respect to such a situation because persons never had the data these tech companies have about them in the first place. Persons were never able to make a longitudinal profile of their eating habits, diversified according to the day in the week, the weather, one's mood and social activities, and was certainly not able to make accurate predictions as to what their food consumption would likely mean for potential medical conditions in 30 years' time. Persons never had the possibility to store their location for every minute of their life, nor had they access to pre-natal data and data about their first years, which are shared with tech companies through the increasingly popular pregnancy apps. And they certainly did not have the capacity to compare their profile to that of thousands or millions of other people. These data are simply produced by tech companies. Thus, the starting position is fundamentally different than in the seventies. Persons used to be the primarily source of information about themselves,

while this has increasingly shifted and is likely to shift even more towards others (such as tech companies).

Second, a new type of problem emerges. A problem typically associated with the affluent availability of data is that others may gain access to private information about a person. The current legal regime addresses this issue by conferring to the individual a right to withhold others access to her private information, home, and body. An equally important, but less theorised, problem is that the individual will be confronted with unwanted information about herself. A photo stored on the web may confront a person with a particular aspect of her past she had long forgotten about (an old love affair; an experimental gothic phase; an anorexic period); personalised content may confront a person with information she was unaware of herself (a Big Data company may infer an adolescent's sexual preference on the basis of the music taste of online friends or infer a person's early pregnancy by analysing her online purchases, while she may not want to find out by being confronted with pregnancy-related advertisements or newsletters that presume a certain sexual preference); data analytics may accurately predict a person's future, while she may rather remain oblivious (eg a FitBit company informing a person that she will presumably suffer from Type 2 Diabetes when she continues her inactive lifestyle). For various reasons, a person may not want to know that she had an anorexic period when she reached puberty, but she may equally not want to be confronted with such information herself, because she wants to forget about that painful period and move on with her life. She may not even mind that her doctor knows that she is suffering from an incurable disease, but still may choose to invoke her right to ask incidental medical findings not to be communicated to her. Persons that want to prevent personal information being communicated to themselves, do not necessarily want to prevent others from accessing or analysing that information.

Third, this creates new problems for the potential to control data and the normative rationale for that control. To start with, why would it be problematic for others to infer data about us, make predictions about our future, or store data about our past that we may want to forget? Why should there be a boundary and where should that boundary be drawn? Perhaps a person does not want to be reminded of the fact that she is obese through weight loss advertisements or that she is likely to get diabetes when she continues her lifestyle, but perhaps it is in her best interest to be reminded. It could even be argued that there is a moral duty for a tech company that has information on a person's likely medical condition to inform her. Again, a problem is the diversity of personal preferences: person A might want to know, person B might not. There is no way of knowing who wants to receive confronting information about themselves and who does not. A final problem for control is that the individual cannot realistically exert it, or she would have to indicate precisely what she does not want to be confronted with, thus being confronted with precisely

that information. Yet it is also highly questionable whether such a moral decision should be left with the data controller or with a DPA.²²

Given what has been said in this chapter, it is difficult to see how the two strategies of data control will work in the twenty-first century. Therefore, what should be considered is to fundamentally reconceptualise the current privacy and data protection framework. What the two approaches have in common is that they focus on controlling data: control is exerted by the individual in the first approach whereas control is exerted by DPAs in the second one. But what either approach fails to appreciate is that control is no longer feasible because of time and resources, but also because of information and power asymmetries: data is produced by data controllers and was thus never in the hands of an individual in the first place. That is why a regime for privacy and data protection that is fit for the twenty-first century should not focus on control over data, but over the means of knowledge production, the technologies, and the resources that are necessary to harvest, infer, and produce data about persons, groups, and society. This could entail regulating technologies and the purposes for which they can be used, this could entail restrictions on the type of data inferences that can be made or this could entail the amount of data and the number of data sources an organisation can have access to. All of these options, obviously, need further research and intense debate before being implemented. Yet while it is not clear which regulatory alternative is best equipped to address the data challenges of the twenty-first century, it is clear that a regime based on control over data is not.

²² Bart van der Sloot, ‘The Right to Be Let Alone by Oneself: Narrative and Identity in a Data-Driven Environment’ (2021) 13(1) *Law, Innovation and Technology* 223.

12

Artificial Intelligence, the Public Space, and the Right to Be Ignored

Andrea Pin

1 Introduction

When they kicked off the endless debate on privacy, Warren and Brandeis complained about the invasion of ‘the sacred precincts of private and domestic life’.¹ They were desperate to insulate individuals from the eyes of private entrepreneurs who took advantage of technological advancements to sell pictures of people caught within private domains.

The right to privacy later morphed into a capacious legal device that—in several legal settings—would protect individuals and groups against private and public powers alike. Technology now allows for detecting, gathering, and processing information about individuals’ identities, behaviours, habits, and characters. Rules in place in certain legal settings, such as the European Union (EU), have severely limited the capacity to gather sensitive personal information such as an individual’s biometrics, thereby limiting what technology can do with people’s identities and personal details.² Such regulations showcase widespread awareness that ‘[a] society that permits the unchecked ascendancy of surveillance infrastructures cannot hope to remain a liberal democracy’.³ What time did not fix, however, was the vulnerability of people in public spaces.

Artificial intelligence (AI)-based technologies can monitor and process information about individuals who find themselves in public places. Several jurisdictions impose constraints on such phenomena, but these are certainly not sufficient. Information gathered in public places is relevant even when individuals themselves are not identified. Technology can detect how fast people walk, guess where they are heading, whether they have something in their hands, or identify what flags they are holding while they protest. Knowing what people hold in their hands

¹ Samuel D Warren and Louis D Brandeis, ‘The Right to Privacy’ (1890) 4(5) *Harvard Law Review* 193, 195.

² Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius, ‘The European Union General Data Protection Regulation: What It Is and What It Means’ (2018) 28(1) *Information and Communications Technology Law* 65–66.

³ Julie E Cohen, ‘What Privacy Is For’ (2013) 126(7) *Harvard Law Review* 1904, 1912.

while they cross a certain square, whether they use sunglasses, or their walking pace, is a valuable piece of information—it may help identify patterns, habits, and tendencies, which inform the economic value of an advertising campaign, or demographics.

The scholarship in the field has devoted its attention to the legal status of public places only rarely.⁴ Constitutional theory has mostly focused on the protection of private places, private lives, and the identity of individuals. Many of those who work in the ever-expanding field of law and tech have also looked elsewhere, focusing on the management, the risks, and the protection of data on the Internet. Especially within the EU, a strong sensitivity to what happens online has emerged:

contemporary data protection law [has been analogised] to environmental regulation. It seeks to protect the democratic ‘commons’, that is, the moral democratic, and cultural environment, as opposed to the natural, physical environment.⁵

Unfortunately, this approach may be misleading: for many tech giants, ‘the key move today is off the Internet’.⁶

The lack of interest in public places is alarming in light of the contemporary widespread recourse of surveillance systems.⁷ It is in public places, in fact, that AI’s capabilities ‘blur the line between being protected and spied on’.⁸ AI-based technologies, new global markets, and law enforcement strategies combine to render public places ideal environments for mass surveillance and monitoring.⁹ Although scholarship cursorily calls to reconceptualise public spaces as shared environments in which individuals and groups enjoy mutual rights and responsibilities,¹⁰ individuals and society at large do not enjoy strong safeguards against this surveillance. Most of the available protection relies on privacy which, unfortunately, ‘is said to be forfeited as soon as a person ventures out in public’.¹¹

⁴ Notable exceptions include: Helen Nissenbaum, ‘Protecting Privacy in an Information Age: The Problem of Privacy in Public’ (1998) 17(5/6) *Law and Philosophy* 559; Elizabeth Paton Simpson, ‘Privacy and the Reasonable Paranoid: The Protection of Privacy in Public Places’ (2000) 50(3) *University of Toronto Law Journal* 305; Sovanartharit Seng and others, ‘A First Look into Users’ Perceptions of Facial Recognition in the Physical World’ (2021) 105 *Computers & Security* 1; Bart van der Sloot, ‘The Right to Be Left Alone by Oneself: Narrative and Identity in a Data-Driven Environment’ (2021) 13(1) *Law, Innovation and Technology* 1.

⁵ Karen Yeung and Lee A Bygrave, ‘Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship’ (2021) 16(1) *Regulation & Governance* 1, 8.

⁶ Amy Kapczynski, ‘The Law of Informational Capitalism’ (2020) 129(5) *Yale Law Journal* 1460, 1470.

⁷ Adam Greenfield, *Everyware. The Dawning Age of Ubiquitous Computing* (New Riders 2006) 107.

⁸ Licia Califano and Valentina Fiorillo, ‘Videosorveglianza’ in R Bifulco and others (eds), *Digesto-Discipline Pubblicistiche* (UTET 2015) 504.

⁹ Roger Brownsword and Han Somsen, ‘Law, Innovation and Technology: Fast Forward to 2021’ (2021) 13(1) *Law, Innovation and Technology* 1, 3 (‘In the Anthropocene, technologies drive the collapse of ... private/public divides’)

¹⁰ Simpson (n 4) 346.

¹¹ ibid 307.

Because of this approach, public places—and places that are publicly observable—have not enjoyed the protection that is accorded to what does not happen in plain sight.¹² The EU may enjoy the reputation of having established a regime of data collection and processing that prioritises human rights, and which many non-EU citizens praise and long for.¹³ This protective approach, however, has not incorporated a significant preoccupation with the transformation that public places are enduring because of AI's power and pressure.

The massive deployment of AI-based technologies in public places around the world is hardly surprising. If, as Shoshana Zuboff maintains, 'companies can stake a claim to people's lives as free raw material for the extraction of behavioral data',¹⁴ public places are the perfect place to gather data, as companies can collect it even from those who are not connected to the Internet. Public spaces are areas within which people share their lives and will unavoidably be exposed to AI-based tools, whether they use them or not.¹⁵ If the new technologies are commonly said to remain in need of a 'guiding philosophy',¹⁶ this is especially true of the status of public places.

This chapter aims to heed these calls for a reconsideration of public spaces and contribute to filling the vacuum in comparative constitutional scholarship. It will proceed by analysing why the topic is so relevant from a social perspective (section 2). Then it will summarily review current contemporary scholarship when it comes to conceptualising public places and the insufficiency of the narrative of privacy in achieving that goal (section 3). Drawing on a variety of legal systems and an array of judicial rulings, it will identify the key legal principles that need to be taken into account in order to devise a more refined understanding of public places that would be helpful in regulating the deployment of AI-based tools (section 4). It will finally make the case for a broader conceptualisation of the issues at stake and for going beyond a purely individualistic approach to the subject (section 5).

2 Reaping the Benefits of AI in Public Spaces

Nowadays, public and private institutions alike are exploiting AI's immense capabilities to monitor and patrol public spaces for a variety of purposes. Just to name a few examples drawn from significantly different environments, China has engineered a comprehensive system of public and private surveillance that connects

¹² Italian Constitutional Court decision no 135 (21 May 2012).

¹³ Yeung and Bygrave (n 5).

¹⁴ Shoshana Zuboff, 'The Coup We Are Not Talking About' *The New York Times* (29 January 2021) <www.nytimes.com/2021/01/29/opinion/sunday/facebook-surveillance-society-technology.html>.

¹⁵ Stamatios Karnouskos, 'Symbiosis with Artificial Intelligence via the Prism of Law, Robots, and Society' (2021) 30(1) *Artificial Intelligence and Law* 1, 1–2.

¹⁶ Henry A Kissinger, Eric Schmidt, and Daniel Huttenlocher, *The Age of AI And Our Human Future* (John Murray 2021) 224.

private devices with video surveillance cameras, which account for 40 per cent of the total number of surveillance cameras in the world.¹⁷ Australia's anti-pandemic policy similarly includes drones that utilise a thermal recognition technology.¹⁸ Moscow has a new facial recognition tool throughout the city,¹⁹ and its schools are integrating a facial recognition system that goes under the telling name of *Orwell*.²⁰ The United States (US)-based *Clearview* Corporation has developed a face recognition software that compares live imaging with pictures of human faces that are already available online through social media and that has been so successful that twenty-four law enforcement agencies outside the US have at least experimented with it.²¹

Of course, such massive deployments of AI-based surveillance among law enforcement agencies spark controversy. Living in an established democracy does not reassure people about these practices, as they perceive themselves as being increasingly monitored, while their faces are restlessly compared with lists of persons of interest, suspects, or criminals.²² After enjoying initial success among law enforcement bodies, *Clearview* has experienced massive backlash and is being rolled back in more than one US state. A recent British case concerning facial recognition software deployed by South Wales Police for two years—and that was able to capture forty images per second—ended up in the Court of Appeal of England and Wales, which found this practice unlawful on a variety of grounds.²³

At least in the West, the concern with AI-based surveillance of public places has focused on one single aspect, namely the identification of individuals. But AI's capabilities do not stop there. Individuals and groups produce crumbs of data of any kind that can be exploited.²⁴ Identifying individuals' and groups' patterns, habits, or trajectories may serve economic or social purposes that have an impact on how citizens live their lives and socialise. Facial analysis systems can measure body temperature, heart rate, or minor facial details such as pupil dilation, drawing inferences about an individual's emotional state.²⁵ Public surveillance systems can

¹⁷ Guglielmo Tamburini, *Etica delle Macchine* (Carocci 2020) 127. A major breakthrough in China was the deployment of mass surveillance technology for the 2008 Beijing Olympics, when it put in place more than two million CCTV cameras in the city of Shenzhen. See Center for AI and Digital Policy, 'AISCI-2020: Facial Recognition Report on Facial Recognition' (February 2020) 5 <file:///C:/Users/utente/Downloads/CAIDP-AISCI-2020-FacialRecognition-(Feb2021).pdf>.

¹⁸ Meredith Van Natta and others, 'The Rise and Regulation of Thermal Facial Recognition Technology during the COVID-19 Pandemic' (2020) 7(1) Journal of Law and the Biosciences 1, 6.

¹⁹ Jonathan Turley, 'Anonymity, Obscurity, and Technology: Reconsidering Privacy in the Age of Biometrics' (2020) 100(6) Boston University Law Review 2179, 2186.

²⁰ Center for AI and Digital Policy (n 17).

²¹ Ryan Mac, Caroline Haskins, and Antonio Pequeno IV, 'Police in At Least 24 Countries Have Used Clearview AI. Find Out which Ones Here' (*BuzzFeed News*, 25 August 2021) <www.buzzfeednews.com/article/ryanmac/clearview-ai-international-search-table>.

²² Cathy O'Neill, *Weapons of Mass Destruction* (Crown Books 2016) 101.

²³ *R (Bridges) v Chief Constable of South Wales Police* [2020] EWCA Civ 1058.

²⁴ Kapczynski (n 6) 1468.

²⁵ Van der Sloot (n 4) 16. On the risks of capturing emotions through public surveillance AI techniques, see Carly Kind, 'Containing the Canary in the AI Coalmine: The EU's Efforts to Regulate

identify and predict walking patterns.²⁶ Private corporations can harvest their knowledge of where people tend to look or stop by while walking, as this may be considered a meaningful detail to estimate what portions of a territory are more economically valuable. Detecting how many people flow into a city may help in designing urban mobility.

Pervasive systems of surveillance do not always equate with Big Brother, of course. In recent years there has been a growing interest in leveraging the capacity of monitoring public places for purposes of utmost importance such as fighting terrorism and the pandemic.²⁷ Monitoring public places may help prevent gatherings that would spread the virus and even identify people who violate quarantine or lockdowns. How far surveillance systems can go in patrolling public places is therefore an extremely timely issue, which does not seem to have received adequate responses so far. Because of the overwhelming importance given to privacy and the Internet, the phenomenon of public places surveillance lacks adequate analysis. The flaws in this approach are reflected in the judicial rulings that often fail to address and even to understand the importance of preserving and protecting not just individuals in public places, but public places themselves, for what they have traditionally offered to citizens.

3 The Problem with Privacy in Public Spaces

Despite their quite different approaches to the topic, China, the US, and the EU share some commonalities. Their legal systems focus either on individuals or private precincts as the two main frameworks for the protection of data: what concerns legal systems seems to be either the protection of individuals or their dwellings. This does not happen by accident: as legal scholarship has largely understood privacy as a shield for the 'liberal self',²⁸ which is centred around the individual in isolation, what happens in places that are accessible to anyone is pushed aside in the theory. Little therefore is said about the protection of public places from the intrusion of surveillance and detection systems. What is public seems to be exposed to public and private powers.

Despite widespread silence on the matter, AI-based surveillance techniques deployed in public spaces challenge the scholars' overwhelming preoccupation with

Biometrics' (*Ada Lovelace Institute*, 30 April 2021) <www.adalovelaceinstitute.org/blog/canary-ai-coalmine-eu-regulate-biometrics/>.

²⁶ Luca Montag and others, 'The Rise and Rise of Biometric Mass Surveillance in the EU' (2021) EJJI 91 <https://edri.org/wp-content/uploads/2021/07/The-Rise-and-Rise-of-Biometric-Mass-Surveillance-in-the-EU_Dutch-Summary.pdf>.

²⁷ Van Natta (n 18) 1.

²⁸ Cohen (n 3) 1905.

AI's impact on privacy and private precincts. Luciano Floridi probably captured the spirit of the time when he coined the idea of *onlife*²⁹—individuals and societies being constantly in flux between the physical world and the online dimension, with their data spilling from the former to the latter without them being aware of it. But the idea of protecting public places is not just about putting the *onlife* dimension under check and limiting the flow of information from and about an individual's physical life to their online dimension. In fact, public places are inhabited by people who may not be—temporarily or permanently—online. Although this is becoming increasingly rare, as more and more people connect via the Internet every day, it is possible to think of individuals and groups not willing to engage in online activities. But since they still live a physical life, peoples' data can be gathered and processed when they are in public places. AI's capabilities can invade physical space and life, as the Pokémon Go saga displayed when millions of individuals around the globe reorganised their movements in order to catch virtual creatures.³⁰ *Pokémon Go* used virtual reality to make the videogame imaginary subjects appear in real locations. Videogame users captured them by taking snapshots of the environments within which such objects suddenly appeared. The process, which caused public disturbance and affected how users navigated public places, thus allowed the corporation to record massive amounts of pictures of private and public places from a variety of angles.

As AI-based systems of surveillance are 'minimally invasive'³¹ cities are the perfect scenario wherein various AI-based mass-monitoring systems can be deployed without imposing physical limitations or restraints on individuals.³² Although people gather, move, and interact with each other on a physical level in public mostly within the urban dimension, political institutions and scholars often overlook this phenomenon, as constitutions and constitutional studies regard cities and urban aggregations as mere parcels of broader territories.³³ AI technologies are therefore radically transforming urban environments, with little or no consideration from constitutional theory.³⁴

Smart cities are a concept that captures how deeply '[m]obile and infrastructural components work together' in surrounding everyone's life³⁵ and silently prompting them to take up new habits. Such changes in how people live in public places may stem from private and public powers alike. Although public institutions can use

²⁹ Luciano Floridi, 'Soft Ethics and the Governance of the Digital' (2018) 31(4) *Philosophy & Technology* 1, 1.

³⁰ Kapczynski (n 6) 1471–72.

³¹ Van Natta (n 18) 7.

³² Joshua A Fairfield, *Owned* (CUP 2017) 62.

³³ Ran Hirschl, *City, State: Constitutionalism and the Megacity* (OUP 2020) 52; Ran Hirschl and Ayelet Schachar, 'Spatial Statism' (2019) 18(1) *ICON* 387, 409.

³⁴ Jathan Sadowski and Frank Pasquale, 'The Spectrum of Control: A Social Theory of the Smart City' (*First Monday*, 2015) <<https://firstmonday.org/ojs/index.php/fm/article/view/5903>>.

³⁵ Fairfield (n 32) 62.

smart-city technologies for public welfare, their massive public surveillance can, nonetheless, provoke significant alterations to individual and social life. Smart cities have thus become a matter of concern,³⁶ as they appear to mimic Bentham's ideal of panoptic surveillance.³⁷

The identification of people who enjoy or cross public places is both one of the most contentious topics and among the most sought after applications of biometric technology.³⁸ Such technology is a matter of contention because public surveillance can have a 'chilling effect'³⁹ on a variety of fundamental freedoms, including 'speech, expression, and association'.⁴⁰ It can discourage individuals from participating in public gatherings or protests for fear of being identified and recorded, as people may be concerned with how public institutions could exploit their data.⁴¹ Surveillance of public places can go as far as looking for individuals who simply jaywalk, or make graffiti.⁴² AI tools can also covertly influence and nudge people in a variety of ways by integrating stealth stimulations in a public environment.⁴³

Although AI surveillance and monitoring systems seem to work passively, as they simply record and process information that is already available, it is actually the bystanders who are passive: they feed AI technologies without being approached and given the opportunity to choose whether they agree to share their data.⁴⁴ The impact of such an imbalance of roles on the social makeup and the political culture has become particularly evident after individuals in the United Kingdom (UK) were fined because they tried to avoid being captured on cameras.⁴⁵ It was their simple reaction to the deployment of AI-based surveillance techniques that discredited them and singled them out as individuals potentially engaging in unlawful activities and disrupting law enforcement procedures.

The deployment of AI tools in public places can disproportionately impact particular social classes, exacerbating economic inequalities. Wealthier individuals or families can expand their privacy by residing in extensive real estate properties, using private transportation, and living more secluded lives.⁴⁶ This would expose

³⁶ ibid 67.

³⁷ Dan L Burk, 'Algorithmic Legal Metrics' (2021) 96(3) *Notre Dame Law Review* 1147, 1155.

³⁸ Turley (n 19) 2213.

³⁹ ibid 2082.

⁴⁰ Simpson (n 4) 342.

⁴¹ Fairfield (n 32) 67.

⁴² Montag (n 26) 22. The Dutch Supreme Court judgment (Hoge Raad App no 02632/02/04, judgment of 20 April 2004) upheld the conviction of a defendant who was caught on camera painting graffiti on the streets, as it found that the deployment of a camera with the purpose of law enforcement did not violate the right to private life as enshrined in art 8 of the European Convention on Human Rights (I am thankful to Evert Stamhuis for pointing out this ruling to me).

⁴³ Van der Sloot (n 4) 18.

⁴⁴ Seng (n 4) 5.

⁴⁵ Turley (n 19) 2196.

⁴⁶ Ari Ezra Waldman, *Industry Unbound* (CUP 2021) 12, sarcastically notes that '[p]rivacy ... has a history of being less of a right for everyone and more of a benefit for the wealthy and privileged.'

them to less surveillance than people of lower socio-economic status, who might live a more exposed social life, as they need to use public transportation, commute more frequently, and share more public facilities or services.⁴⁷ The legal treatment of such technologies is, therefore, critical to preserve everyone's equal capacity to develop and enjoy social relationships,⁴⁸ regardless of social strata.

4 Going Beyond Privacy

Legal systems are clearly going in different directions in terms of promoting or limiting the deployment of surveillance technology. China envisions a near future within which it will be able to almost immediately identify any one of its 1.3 billion citizens.⁴⁹ Several US states are going the other way, banning or limiting the usage of drones, body cameras, and other tools previously deployed to monitor public places and identify individuals through AI-based surveillance techniques.⁵⁰ The EU famously protects privacy in the form of personal data in the public sphere and is anticipating discouraging AI-based surveillance that deploys facial recognition systems as a high-risk practice for the sake of such data protection.⁵¹ The AI Bill that is under consideration by EU institutions certainly considers public surveillance as a potentially high-risk practice, but does not rule it out nor does it seem to really consider factors that go beyond the identification of individuals. It therefore still omits a thorough conceptualisation of what surveillance of public places entails for law and society at large.

Articulating the need to preserve a level of privacy and discretion for individuals in public places has been quite challenging. As someone has noted, what is at stake within this scenario seems to be different from 'conventional privacy'.⁵² Privacy, in fact, is usually believed to protect 'people, not places'.⁵³ And that is where the problem begins.

As the US Supreme Court has stated, the expectation of privacy 'generally ends when one enters the public'.⁵⁴ The possibility of being observed by the naked eye dissolves such an expectation, in the US as elsewhere.⁵⁵ The often implicit assumption underpinning this approach asserts that people cannot reasonably expect to enjoy privacy while they socialise or occupy shared places, so they should simply

⁴⁷ Simpson (n 4) 345.

⁴⁸ Eoin Carolan, 'Stars of Citizen CCTV: Video Surveillance and the Right to Privacy in Public Places' (2006) 13(1) Dublin University Law Journal 326, 345.

⁴⁹ Turley (n 19) 2185.

⁵⁰ *ibid* 2194.

⁵¹ See the chapter by Natalia Menéndez González in this volume.

⁵² Turley (n 19) 2196.

⁵³ *Katz v United States* 389 US 347, 351 (1967).

⁵⁴ Turley (n 19) 2202.

⁵⁵ *ibid* 2203. See also Simpson (n 4) 306.

avoid doing in public what they prefer not to share with others.⁵⁶ In other words, when an individual is in public, it is her responsibility to keep information about herself discreet.

This assumption, however, has meant relinquishing most safeguards whenever an event or a person is observable with the naked eye. Even surveillance from above—which is not exactly what ‘reasonable people’ would expect—has been affirmed as lawful when it covered public spaces.⁵⁷ In fact, even the English and Welsh Court of Appeal in *Bridges*—the very judgment that found the deployment by South Wales Police of face recognition technology to be unlawful in light of EU law—did not rule out this possibility. It simply found that the deployment itself left too much discretionary power to the police in selecting the individuals that it was attempting to locate and further found that the software that had been deployed could embed discriminatory biases.

The problem with privacy is how a century of scholarship, judicial rulings, and legislative activities have construed it. It is a capacious and multifaceted term, as it includes both the ‘physical and psychological integrity’ of a person, thereby embracing ‘multiple aspects of the person’s physical and social identity’.⁵⁸ However, it seems to end where the public begins. Despite significant historical differences, legal traditions as diverse as those of Italy,⁵⁹ Ireland,⁶⁰ New Zealand,⁶¹ the UK,⁶² and the US—just to name a few—seem to agree on one point: privacy claims in public places are hardly defensible because what meets the eye obviously cannot be said to be private, or to be believed to be private. This approach deprives even facts or details that were visible only under very special circumstances of protection: the Irish High Court found that events taking place within private precincts did not enjoy privacy if they were observable with the help of a ladder,⁶³ while the US Supreme Court stated that what a helicopter could notice within private properties did not infringe privacy expectations.⁶⁴

The average theorisation of what a privacy violation would consist of also weakens the articulation of privacy to cover public places. The concept of privacy usually associates the infringement with an act of ‘intrusion’⁶⁵ or ‘invasion’⁶⁶—those violating someone’s privacy are perceived as ‘intruding’ upon their sacred precincts. Nothing of this kind occurs in public.

⁵⁶ Simpson (n 4) 321.

⁵⁷ *California v Ciraolo* 476 US 207 (1986) [213].

⁵⁸ *Marper v the United Kingdom* App nos 30562/04 and 30566/04 (ECtHR, 4 December 2008), para 66.

⁵⁹ Italian Constitutional Court decision no 149 (16 May 2008).

⁶⁰ *Atherton v Director of Public Prosecutions* [2005] IEHC 429.

⁶¹ *Hosking v Runiting Ltd* [2004] NZCA 101-03.

⁶² *Campbell v MGN Ltd* [2004] UKHL 22, [2004] 2 All ER 995 (Lord Nicholls of Birkenhead).

⁶³ *Atherton* (n 60) [19].

⁶⁴ *Florida v Riley* 488 US 445 (1989).

⁶⁵ Nissenbaum (n 4) 569.

⁶⁶ *Hosking* (n 61) 246 (Tipping).

Admittedly, scholarship has occasionally urged to reconceive and expand privacy to cover information gathered in public places,⁶⁷ especially after AI-based surveillance systems took off. Scholars have become increasingly aware that such technologies are altering the self-perception of individuals whenever they find themselves in public places.⁶⁸

It is common sense that social life requires that individuals give up their expectation of enjoying the same degree of privacy that they expect in private.⁶⁹ When people are in public they are vulnerable, in the sense that they can be observed and noticed. However, there is a general expectation that such an event is unlikely to happen. People, in fact, do not imagine ‘that they are being observed any more than casually and by a limited number of people’.⁷⁰ But this is not the case anymore with AI-based surveillance. Such technologies can notice an individual’s unique features on a large scale: for instance, the automated facial recognition system deployed by the South Wales police was able to scan forty faces per second.⁷¹ Moreover, events and people that take place or appear in public can be recorded and processed; recording an event or a detail ‘is different in kind, not merely in degree, from being able to relate it verbally or even by way of a sketch’.⁷² Being noticed and remembered, which would usually happen by accident and cursorily, becomes the standard.

The European Court of Human Rights (ECtHR) has already tried to expand the protection of privacy in public places in the context of surveillance. Although admitting that ‘[a] person who walks down the street will, inevitably, be visible to any member of the public who is also present’, it has nonetheless recognised that technological monitoring of an area may raise privacy issues ‘once any systematic or permanent record comes into existence of such material from the public domain’.⁷³

However, the need for a sort of privacy in public looks substantially different from what one would expect when she is in private. If we look for the interest in not being spied on while in public, we would probably find a better definition of it in the concept of ‘*anonymity*’,⁷⁴ namely the unlikelihood that an event or a detail is connected with a specific individual or a group. Such anonymity is critical, as it enables people to socialise and participate in initiatives and events of public significance, such as protests, without fear of retaliation or other types of adverse consequences.⁷⁵ More broadly, it preserves people and their habits from being put

⁶⁷ Nissenbaum (n 4) 559.

⁶⁸ Simpson (n 4) 340.

⁶⁹ *Hosking* (n 61) 125 (Gault).

⁷⁰ Simpson (n 4) 321.

⁷¹ *Bridges* (n 23) 16.

⁷² *Creation Records Ltd v News Group Newspapers Ltd* [1997] EWHC 370 (Ch) [29] (Lloyd).

⁷³ *PG v United Kingdom* App no 44787/98 (ECtHR, 25 September 2001), para 57.

⁷⁴ Simpson (n 4) 325.

⁷⁵ Turley (n 19) 2196.

under a spotlight. Anonymity serves as a ‘breathing space’ for individuals who find themselves *in* a public place.⁷⁶ It is, however, possible to conceptualise also the anonymity of public places: what happens in publicly accessible areas is not expected to be processed and remembered. Public places have thus provided people with breathing spaces, but also society at large with areas of socialisation precisely because of the expectation that people and social relations enjoy anonymity within them.

AI-based surveillance undermines anonymity *in* and *of* public places. Anyone can be recognised by chance—or a certain event can be spotted and memorised by bypassers or bystanders. As has been contented, ‘[m]ost people reasonably make this assumption: either that they are not noticed, or that any single observer can observe and harbor only discrete bits of information’.⁷⁷ Public spaces’ surveillance systems that use AI make certain what should be unlikely. They spot, identify, and record events. They can gather and process them, making inferences that may have a huge-scale impact on those who are in those places. In fact, AI allows for the gathering, recording, storage, processing, and analysis of a potentially unlimited amount of information.⁷⁸

This is where the concept of ‘*confidentiality*’ may usefully kick in and supplement the legal analysis with additional factors. Courts around the world have more than once danced between confidentiality and privacy, sometimes arguing that the two concepts capture the same essence.⁷⁹ As Lord Phillips has noted, however, there is a gap between these two concepts. Confidentiality requires that someone is informed about someone else’s sensitive information, and a breach of confidentiality happens when the information is spread to other people.⁸⁰

Strictly speaking, this concept of confidentiality certainly does not respond to the challenge of surveillance in public places. Those who walk a certain square do not willingly share their information with the cameras and other devices patrolling a certain area, and such sharing does not establish a duty of confidence.⁸¹ But they still do share information, which may bear consequences, even at the aggregate level. The information about the shopping behaviours, ideological inclinations, or habits of a certain portion of the population can affect people’s lives, as they can become targets of economic or political campaigns.

As some judicial rulings have acknowledged, a capacious concept of confidentiality helps at least in understanding that there are boundaries regarding what can be done with such data. In fact, as the Court of Appeal of England and Wales has

⁷⁶ *ibid* 2228.

⁷⁷ Nissenbaum (n 4) 576.

⁷⁸ *ibid* 576.

⁷⁹ *Douglas v Hello! Ltd* [2005] EWCA Civ 595 [83]: ‘for the adjective “confidential” one can substitute the word “private”’ (Lord Phillips).

⁸⁰ *ibid* 55 (Lord Phillips).

⁸¹ *Hosking* (n 61) 26 (Gault).

stated, '[a] duty of confidence will arise whenever the party subject to the duty is in a situation where he knows or ought to know that the other person can reasonably expect his privacy to be protected'.⁸² Or, in the words of New Zealand's Court of Appeal, 'a pre-existing confidential relationship between the parties is not required for a breach of confidence claim. The nature of the subject matter or the circumstances ... may suffice to give rise to liability'.⁸³ Similarly, a sort of mutual expectation of confidentiality among people in public places who most of the time do not know each other, may be said to generate the duty not to further disclose and utilise the information.

The England and Wales's and the New Zealand's court statements help understand how public places should be treated. If people have an expectation of anonymity, then what is captured by surveillance systems might be seen as a sort of confidential information. And, as AI-based surveillance technology can gather crumbs of information to infer what types of individuals are walking down a certain street or where they are heading, the 'subject matter' covered by confidentiality may include bits of information that may appear to be without value, but the aggregation of which may be economically and politically valuable.

Attempts to tinker with the concept of privacy and make it broad enough to encompass what happens to the individual in public seem to be fraught with contradictions. At best, it limits what surveillance systems can capture, record, and process, but not the phenomenon of surveillance in itself.⁸⁴ Looking at the case law, it is quite difficult to square the circle and provide individuals with privacy in public places. Such a sea change requires going beyond the ideas of private precincts, intrusion, invasion, and observability.

What the scholarship that has been trying to expand the notion of privacy to cover public places certainly got right is that current technologies are different from the occasional chance of someone spotting someone else that is known to her in the crowd, or a specific detail of an individual. The difference is not a matter of degree: it is rather a matter of quality. In fact, it changes the character of public places, depriving them of their anonymity.

Anonymity may not protect individuals and events that happen in public, as they are both observable and noticeable. But this is where confidentiality makes a contribution to the subject, especially if—as some have maintained—confidentiality may arise not just from a contract, but also from someone's expectations in light of the specific circumstances. Although it is construed as an obligation among

⁸² *Douglas* (n 79) 72 (Lord Phillips).

⁸³ *Hosking* (n 61) 33 (Gault).

⁸⁴ In the field of personal identity recognition, China's new regulation on data protection also protects personal data, unless there is a public interest involved, and therefore echoes the privacy-oriented Western approach. See Editorial, 'China's New Comprehensive Data Protection Law: Context, Stated Objectives, Key Provisions' (*Future of Privacy Forum*, August 2020) <<https://fpf.org/blog/chinas-new-comprehensive-data-protection-law-context-stated-objectives-key-provisions/>>.

specific persons, confidentiality certainly fosters a sense of discretion, conveying the need that what is seen is not spread and exploited.

5 Going Beyond Individualism in Public Places

Protecting public places from boundless exploitation of information regarding people gathered through pervasive AI surveillance and monitoring technologies does not go simply beyond the idea of privacy. It also goes beyond the idea that the individual is the only subject worthy of legal protection. This is another important step that is worthwhile considering in more depth.

Privacy protects *personal information*. Most jurisdictions protect certain information because it belongs to individuals. As long as information is disconnected from individuals, it loses much of its protection. The same is true about confidentiality.

AI-based surveillance is not problematic just because it identifies individuals and singles out specific details about them. Its impact is much broader than that. In fact, it can identify patterns, average age, walking speed, where people tend to go, where they usually slow down, and what they hold in their hands.

This type of information is of great importance for private businesses and public institutions alike. The speed of pedestrians can help estimate the worth of a billboard; their average age can supply information about the best shops to locate in a certain area; the walking patterns can help coordinate urban traffic and planning; from the way people dress AI can infer their wealth and cultural and political leaning; their skin colour can help determine if some urban areas are de facto racially segregated.

Such information may not just be of great value. It can also pose grave threats to the well-being of residents. Urban policies and private corporations can use this information to target selected strata of the population and discourage, encourage, or nudge certain economic classes into specific behaviours. They can quietly manipulate urban societies.

The threat is grave also when the population is conscious that AI-based surveillance tools are patrolling public areas. Public places are both areas of anonymity as well as places of socialisation among peers; the importance of such areas for people is undeniable. The fact that their activities are spied on, recorded, processed, and gathered seems to also change the role of such places. The fact that people know they are being monitored may deter them from walking in public places and therefore boost the ‘atomization of social life’⁸⁵ that is already underway: individuals

⁸⁵ Angioletta Sperti, ‘The Impact of Information and Communication Revolution on Constitutional Courts’ in Martin Belov (ed), *The IT Revolution and Its Impact on State, Constitutionalism and Public Law* (Bloomsbury 2021) 184.

may feel safer at home and prefer socialising online rather than offline. The relationship between individuals and monitoring and surveillance systems in public thus renders the necessity of protecting public places particularly acute. These aspects urge to reframe the issue of technological surveillance of public places and to go beyond the privacy paradigm and the focus on the individual.

6 Conclusion

Despite some scholarly efforts, constitutional theories generally have failed to address the uniqueness of public places and their relationship to privacy. Scholars who are aware of the perils of uninterrupted data mining on individuals suggest that '[w]e need consciously to create privacy zones in order to claw back some areas in which creativity and freedom can take flight unimpeded'—they therefore theorise that some areas should be preserved as secluded from the public.⁸⁶ The very idea of reaffirming the private/public distinction and the dichotomy between individuals and the public⁸⁷ as a viable solution, however, does not seem to grasp the meaning and uniqueness of public places for private and public institutions alike. The overwhelming focus on individuals also overlooks the fact that surveillance technologies have an impact on societies at large, not just on persons taken in isolation. As Virginia Eubanks has it, '[i]n his famous novel 1984, George Orwell got one thing wrong. Big Brother is not watching *you*, he's watching *us*'.⁸⁸

Both individuals and social life as a whole need protection from the ever-growing phenomenon of monitoring and surveillance. Courts have certainly identified safeguards by leveraging privacy rights, and legislatures are taking action around the world against massive deployments of such technologies. It is the notion of public places itself, however, that is lacking a proper assessment. This gap leaves people who are even offline in physical public places less protected than people who are online but can avail themselves of more crafted forms of legal protection that revolve around the concepts of personal data and privacy.

Drawing from existing case law in some jurisdictions that have shown sensitivity to the concepts of privacy in public places, I have proposed to tap into the issue of AI-based surveillance of public places the notions of anonymity and confidentiality. These concepts were conceived when information about *specific people and details of their lives were caught and spread by specific individuals and businesses*. But they can be married together to identify information that should not be shared and exploited *because of the circumstance* within which they are caught. In other words, also information gathered from AI-based tools in public places should also

⁸⁶ Carissa Véliz, *Privacy Is Power* (Melville House 2021) 195.

⁸⁷ Antoine Garapon and Jean Lassègue, *Justice Digitale* (PUF 2018) 86–87.

⁸⁸ Emphasis in original. Virginia Eubanks, *Automating Inequality* (St Martin 2018) 6.

be covered by some degree of anonymity and confidentiality. This would not just protect individuals. It would also protect social life, thereby creating a general rule that prohibits public surveillance, except for reasons that are valuable for society at large. After all, some have already argued that privacy is nothing else than ‘a facet of social life that gives people the confidence and moral space to share information with others’.⁸⁹ Vindicating this aspect of privacy is easier if privacy is married with anonymity and confidentiality. This would re-balance the equilibrium between the vast population whose information is captured inadvertently and discretely, and the private and public institutions that can exploit such information.

⁸⁹ Waldman (n 46) 51.

PART IV

ARTIFICIAL INTELLIGENCE AND
NON-DISCRIMINATION

13

Artificial Intelligence and Racial Discrimination

Louis Koen and Kgomo Mufamadi

1 Introduction

The rise in the use of artificial intelligence (AI) as an integral part of everyday decision-making has contributed to growing scholarly interest in the human rights implications of these technologies. This includes an increased focus on racial discrimination in the application of these technologies and the extent to which these technologies entrench existing inequalities.¹ Strong evidence has emerged indicating the racially discriminatory effect of various types of AI technology. The nature of this discrimination ranges from direct racial discrimination ‘explicitly motivated by intolerance or prejudice’ to less explicit forms of indirect discrimination which appear neutral but have a disproportionate impact on certain racial groups.²

This chapter provides an overview of contemporary concerns over racial discrimination in the application of AI before critically analysing the contemporary international law framework for combatting racial discrimination. It includes an analysis of states’ due diligence obligations, whilst noting that most AI applications are designed and developed by private entities. It then proceeds to consider the value of imposing positive obligations on government and private entities alike to prohibit racial discrimination and take active steps to eliminate such discrimination.

2 The Prevalence of Racial Discrimination in AI

What is clear from the literature is that the concern over AI having discriminatory outcomes is not new. It has been suggested that AI is characterised by slow

¹ Frederik J Zuiderveen Borgesius, ‘Strengthening Legal Protection Against Discrimination by Algorithms and Artificial Intelligence’ (2020) 24 International Journal of Human Rights 1572.

² Office of the United Nations High Commissioner for Human Rights (OHCHR), ‘Report of the Special Rapporteur on Racial Discrimination and Emerging Digital Technologies: A Human Rights Analysis’ (2020) UN Doc A/HRC/44/57.

developments followed by sudden great leaps forward. Yet, bias subtly permeates this work in a manner that tech companies are hesitant to acknowledge.³ Additionally, it has been noted that AI is ultimately created by humans and used in systems and institutions that have notoriously entrenched discrimination, ranging from the criminal justice system to housing, the workplace, and our financial systems.⁴ Frequently, the data used to train AI contains bias. Belenguer notes that the data is not only biased, but also unrepresentative of people of colour, women, and other marginalised groups.⁵ The challenges around representation in tech industries exacerbate the problem.⁶ Researchers highlight numerous examples of this. For example, AI in a Google online photo service organised photos of Black people into a folder called 'gorillas'.⁷

In relation to the criminal justice system, the United Nations (UN) News website comments on a case involving an African American man in the United States (US) who was arrested for shoplifting after the police officers involved relied on facial recognition AI. However, the tool relied upon had not learned how to recognise the differences between the faces of Black people because the faces used to train it had mostly been white.⁸ The article notes further that there is considerable evidence that AI is making the world more unequal and benefits a small proportion of people; it adds:

For example, more than three-quarters of all new digital innovations and patents are produced by just 200 firms. Out of the fifteen biggest digital platforms we use, eleven are from the US, whilst the rest are Chinese. This means that AI tools are mainly designed by developers in the West. In fact, these developers are overwhelmingly white men, who also account for the vast majority of authors on AI topics.⁹

³ Tugrul Kisken, *Towards an International Political Economy of Artificial Intelligence* (Springer Nature 2022) 15.

⁴ Lindsay Weinberg, 'Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches' (2022) 74 *Journal of Artificial Intelligence Research* 75.

⁵ Lorenzo Belenguer, 'AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry' (2022) 2 *AI and Ethics* 771.

⁶ *ibid.*

⁷ Cade Metz, 'Who Is Making Sure the AI Machines Aren't Racist?' *The New York Times* (15 March 2021) <www.nytimes.com/2021/03/15/technology/artificial-intelligence-google-bias.html>.

⁸ UN News, 'Bias, Racism and Lies: Facing up to the Unwanted Consequences of AI' (UN News, 30 December 2020) <<https://news.un.org/en/story/2020/12/1080192>>. For a general discussion on error in facial recognition software, see also Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (First Conference on Fairness, Accountability and Transparency, Nice, 2018) <<http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>>.

⁹ UN News (n 8).

In the criminal justice system, judges in the US have used the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software tool to evaluate the likelihood of an individual reoffending and, consequently, what an appropriate sentence would be for the particular individual. The technology is meant to reduce bias in law enforcement, but instead, it reinforces stereotypes of people of colour being repeat offenders.¹⁰ Several studies show that it misinterprets Black people as high-risk reoffenders twice as often as it does white people.¹¹

The challenges around discriminatory AI also came to the fore during the COVID-19 pandemic in relation to critical care resource allocation in the US. One of the so-called benefits many proponents of AI highlight is that AI is ‘colour-blind’ and ought to, therefore, eliminate racism in certain systems. However, Williams and others¹² have analysed a system of critical care resource allocation developed by White and others, showing otherwise.¹³ While the system does not incorporate race in its decision-making algorithm, its use ensures that white people receive disproportionately more critical care resources than people of colour.¹⁴ This is the case because the system is based on a scoring system which uses pre-existing data to determine the need for critical care resources such as life expectancy. In the US, Black people have an increased risk of high blood pressure, kidney disease, diabetes, and certain cancers, inevitably lowering their life expectancy.¹⁵ Accordingly, if this data is used to determine COVID critical care resources it will inevitably lead to Black communities receiving broadly less critical resources. Its continued use beyond the pandemic would also see these problems in the allocation of healthcare resources apply more broadly.¹⁶

Furthermore, AI risks perpetuating housing discrimination such as tenant selection and mortgage qualifications. It has been observed that despite their ability to pay rent, people are frequently denied housing because tenant-screening algorithms find them unfit or unacceptable.¹⁷

¹⁰ Lel Jones, ‘A Philosophical Analysis of AI and Racism’ (2020) 13 *Stance* 36, 41.

¹¹ Melissa Hamilton, ‘Investigating Algorithmic Risk And Race’ (2021) 5 *UCLA Criminal Justice Law Review* 530.

¹² J Corey Williams and others, ‘Colourblind Algorithms: Racism in the Era of COVID 19’ (2020) 112 *Journal of the National Medical Association* 550.

¹³ DB White and authors, ‘Allocation of Scarce Critical Care Resources during Public Health Emergency’ (Department of Critical Care Medicine, University of Pittsburgh School of Medicine, 2020) <<https://bioethics.pitt.edu/sites/default/files/Univ%20Pittsburgh%20-%20Allocation%20of%20Scarce%20Critical%20Care%20Resources%20During%20a%20Public%20Health%20Emergency.pdf>>.

¹⁴ *ibid.*

¹⁵ BW Ward and JS Schiller, ‘Prevalence of Multiple Chronic Conditions Among US Adults: Estimates from the National Health Interview Survey, 2010’ (2013) 10 *Preventing Chronic Disease* E65 <<https://doi.org/10.5888/pcd10.120203>>.

¹⁶ Michael Rigby, ‘Ethical Dimensions of Using Artificial Intelligence in Health Care’ (2019) 21 *AMA Journal of Ethics* 122.

¹⁷ Olga Akselrod, ‘How Artificial Intelligence Can Deepen Racial and Economic Inequities’ (ACLU, 13 July 2021).

Jones notes that AI is also absorbing job functions which are disproportionately performed by Black and Latinx people in the US and being redistributed to the tech industry,¹⁸ where only 7.5 per cent of workers are Black and Latinx.¹⁹ Importantly, it is not the machines that are responsible for discrimination in AI, but the human beings behind the AI, the methods used to code the AI, and the institutions within which they operate. And although there are some benefits to AI, it is arguably unjustifiable that efficiency is prioritised over justice to the detriment of marginalised groups of people. As Jones correctly notes, the most critical work of the tech industry to address these concerns must inevitably involve more proportional representation in the tech industry and software data samples which include data from all races and ethnicities.²⁰

3 International Law and the Obligation to Prevent Racial Discrimination

On 21 December 1965, the UN General Assembly (UNGA) unanimously adopted and opened for signature the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD). The adoption of the ICERD had been preceded, as noted by Schwelb, by an outburst of swastika painting and 'other manifestations of anti-Semitism and other forms of racial and national hatred and religious and racial prejudices of a similar nature' across many countries in the winter of 1959–60.²¹ Due to these racially and religiously motivated hatred incidences the Sub-Commission on Prevention of Discrimination and Protection of Minorities, which had been in session at the time, saw fit to adopt a resolution condemning these acts and started taking the initiative to collect information about these events with the ultimate aim of recommending corrective action.²²

The proposed action was approved by the Sub-Commission's superior bodies, including the UNGA.²³ When the Sub-Commission evaluated the data gathered at its 1961 session, it was suggested that the UNGA be encouraged to prepare an international convention imposing specific legal obligations on the parties to prohibit manifestations of racial and national hatred.²⁴ What was considered at the time was a far narrower instrument than what eventually became the ICERD.²⁵ The

¹⁸ Jones (n 10).

¹⁹ US Equal Employment Opportunity Commission, 'Diversity in High Tech' <www.eeoc.gov/special-report/diversity-high-tech>.

²⁰ Jones (n 10) 45.

²¹ Egon Schwelb, 'The International Convention on the Elimination of all Forms of Racial Discrimination' (1966) 15 International and Comparative Law Quarterly 996.

²² *ibid* 997.

²³ Theodor Meron, 'The Meaning and Reach of the International Convention on the Elimination of all Forms of Racial Discrimination' (1985) 79 American Journal of International Law 283.

²⁴ Patrick Thornberry, *The International Convention on the Elimination of All Forms of Racial Discrimination* (OUP 2016) 25.

²⁵ Schwelb (n 21) 997.

ICERD represented the first international instrument specifically targeted at the elimination of racial discrimination and now has 182 state parties.

The ICERD defines racial discrimination to mean:

[A]ny distinction, exclusion, restriction or preference based on race, colour, descent, or national or ethnic origin which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life.²⁶

This definition is relatively broad, with Judge Iwasawa of the International Court of Justice (ICJ) noting that the establishment of racial discrimination is subject to two conditions: The first need is that there must be a preference, exclusion, restriction, or differentiation 'based on race, colour, descent, or national or ethnic origin'. Second, the differentiation must have the 'goal or effect' of undermining or preventing the equal recognition, enjoyment, and exercise of human rights and basic freedoms in all spheres of public life, including politics, the economy, society, culture, and other areas.²⁷

The usage of the term 'effect' indicates that the ICERD definition of racial discrimination also covers measures resulting in indirect discrimination. Former ICJ Judge Crawford commented on this in noting that a law, measure, or restriction may be considered racial discrimination if it has the 'effect' of restricting the exercise or enjoyment of the rights enshrined in the ICERD on an equal basis.²⁸ The Committee on the Elimination of Racial Discrimination (CERD) has similarly observed that:

[T]he definition of racial discrimination in article 1 expressly extends beyond measures which are explicitly discriminatory, to encompass measures which are not discriminatory at face value but are discriminatory in fact and effect, that is, if they amount to indirect discrimination. In assessing such indirect discrimination, the Committee must take full account of the particular context and circumstances of the petition, as by definition indirect discrimination can only be demonstrated circumstantially.²⁹

²⁶ ICERD, art 1(1).

²⁷ *Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Qatar v United Arab Emirates)* (Preliminary Objections, Judgment of 4 February 2021) [2001] ICJ Rep 71 [48] (dissenting opinion of Judge Iwasawa).

²⁸ *Application of the International Convention for the Suppression of the Financing of Terrorism and of the International Convention on the Elimination of All Forms of Racial Discrimination (Ukraine v Russian Federation)* (Provisional Measures, Order of 19 April 2017) [2017] ICJ Rep 215 [7] (declaration of Judge Crawford).

²⁹ *LR v Slovakia* Communication No 31/2003 (CERD Committee, 7 March 2005), para 10.4.

The need to regulate both direct and indirect forms of racial discrimination is particularly important in the context of AI. In many instances, discrimination arises from algorithms which may not have been deliberately programmed to discriminate.³⁰ However, the data used in the algorithm may display existing bias and in so doing perpetuate or worsen existing forms of discrimination against people of colour.³¹ This has proven to be particularly problematic in instances where Black communities had traditionally been over-policed and saw a greater prevalence of crime in statistics because of this over-policing. Predictive crime algorithms would then perpetuate these stereotypes as it is built on bad data.³² Therefore, while the algorithm does not necessarily discriminate intentionally, the outcomes based on discriminatory data will disproportionately negatively affect communities of colour.³³

Article 2 of the ICERD imposes direct obligations on state parties to combat racism including by not engaging in racial discrimination,³⁴ as well as to refrain from sponsoring, defending, or supporting racial discrimination by any person or organisation,³⁵ and an obligation to prohibit discrimination by any person or organisation.³⁶ The obligation to prohibit discrimination by any person or organisation requires the state to explore all appropriate means, including adopting legislation to eradicate racial discrimination. This means that the state has a clear obligation to prevent racial discrimination even where the perpetrator of the discrimination is a private party such as an AI manufacturer.

Article 5 of the ICERD goes on to expand on the state's obligation to prohibit and eliminate racial discrimination in the enjoyment of various rights such as equal treatment before the tribunals and all other organs administering justice,³⁷ and the rights to work, to free choice of employment, to just and favourable conditions of work, to protection against unemployment, to equal pay for equal work, to just and favourable remuneration.³⁸ The obligation to eliminate racial discrimination is not merely a negative obligation imposed on the state not to discriminate but requires the state to take positive measures towards the attainment of these objectives.

The importance of a positive duty to eliminate discrimination in the context of AI was aptly demonstrated in the case of *R (Bridges) v South Wales Police*,³⁹ decided by the England and Wales Court of Appeals. The case concerned the legality of the South Wales Police (SWP) Force's use of live automated facial recognition (AFR)

³⁰ Borgesius (n 1) 1577.

³¹ *ibid.*

³² OHCHR, 'Report of the Special Rapporteur on Visit to the United Kingdom of Great Britain and Northern Ireland' (2019) UN Doc A/HRC/41/54/Add.2, para 40.

³³ *ibid.*

³⁴ ICERD, art 2(1)(a).

³⁵ *ibid* art 2(1)(b).

³⁶ *ibid* art 2(1)(d).

³⁷ *ibid* art 5(a).

³⁸ *ibid* art 5(e)(i).

³⁹ *R (Bridges) v South Wales Police* [2020] EWCA Civ 1058 (11 August 2020).

technology in an ongoing trial using a system called ‘AFR Locate’. AFR Locate entails the deployment of surveillance cameras to capture digital images of members of the public, which are then processed and compared with digital images of people on a watch list compiled by the SWP.⁴⁰ The system was used to identify various categories of persons including people who are wanted on a warrant and people who are simply of interest to the police for intelligence purposes.⁴¹

A human rights group then challenged the use of this AI application in the Divisional Court of the Queen’s Bench Division (Divisional Court).⁴² The Divisional Court rejected the claim that the SWP failed to comply with its obligation to combat discrimination by not considering the possibility that AFR Locate might produce results that were indirectly discriminatory on the basis of sex and/or race because it produces a higher rate of positive matches for female faces and/or Black and ethnic minority faces.⁴³ According to the Divisional Court, there had been no evidence that the AI application in question produced any such results.⁴⁴ However, this decision was overturned on appeal, with the Court of Appeal noting that the question of evidence as to discrimination:

[W]as to put the cart before the horse. The whole purpose of the positive duty (as opposed to the negative duties in the Equality Act 2010) is to ensure that a public authority does not inadvertently overlook information which it should take into account.⁴⁵

The appropriate question before the court on review was accordingly not whether the adoption of the AI application has actually discriminated on the basis of race or gender.⁴⁶ Instead, the positive duty to eliminate and prevent racial discrimination requires a public authority to show that it has taken active steps to evaluate the risk of software having a potential racial bias.⁴⁷ This positive duty was sourced from section 149(1) of the Equality Act 2010 (EA 2010) which provides that public authorities must have due regard to the need to ‘eliminate discrimination, harassment, victimisation and any other conduct that is prohibited’ in terms of the Act. The court held that the mere potential for racial bias in itself requires the undertaking of due diligence to prevent the discrimination from ever occurring rather than requiring a litigant to prove the discrimination after it has already occurred.⁴⁸ The use of AFR Locate was held to have violated the EA 2010 on account of the

⁴⁰ *ibid* [1].

⁴¹ *ibid* [13].

⁴² *R (Bridges) v Chief Constable of South Wales Police* [2019] EWHC 2341 (Admin).

⁴³ *ibid* [153].

⁴⁴ *ibid* [153]–[158].

⁴⁵ *R (Bridges) v South Wales Police* (n 39) [182].

⁴⁶ *ibid* [182].

⁴⁷ *ibid* [201].

⁴⁸ *ibid* [201].

state's failure to conduct a proper risk assessment and the implementation of measures to prevent discrimination.⁴⁹

It is submitted that the positive duty to eliminate discrimination in the ICERD similarly requires states to implement measures to prevent discrimination from ever occurring. It is also not unreasonable to expect state parties to undertake appropriate risk assessments before adopting potentially discriminatory AI. Indeed, it is better to address the risk of bias before any adverse consequences arise, or as the court noted:

We would hope that, as AFR is a novel and controversial technology, all police forces that intend to use it in the future would wish to satisfy themselves that everything reasonable which could be done had been done in order to make sure that the software used does not have a racial or gender bias.⁵⁰

Similarly, this obligation to undertake an appropriate risk assessment should not be limited to the state. Given that the ICERD also imposes an obligation on the state to prevent racial discrimination by private parties, the state is clearly required to have legislation in place requiring companies and other private entities to eliminate discrimination.

4 Business Due Diligence in the Prevention of Racial Discrimination and Positive Duties to Eliminate Racial Discrimination

Within the context of equality law, scholars have increasingly considered the limitations of a focus on negative obligations such as the obligation not to discriminate *per se*.⁵¹ For instance, Sturm has argued that reducing more subtle forms of discrimination requires a structural approach.⁵² According to Sturm, the structuralism approach requires a regulatory strategy that supports the establishment of institutions and mechanisms to implement general principles in specific contexts.⁵³ Harpur, moreover, refers to the focus on positive duties as a management-based approach and argues that, in terms of this approach, equity issues are not addressed after a system has been built, but rather throughout the conception,

⁴⁹ *ibid* [202].

⁵⁰ *ibid* [201].

⁵¹ Frank Dobbin and Alexandra Kalev, 'Multi-Disciplinary Responses to Susan Sturm's The Architecture of Inclusion: Evidence from Corporate Diversity Programs' (2007) 30 Harvard Journal of Law and Gender 279.

⁵² Susan Sturm, 'Second Generation Employment Discrimination: A Structural Approach' (2001) 101 Columbia Law Review 458.

⁵³ *ibid*.

implementation, and operation of the system. Consequently, duty bearers are more likely to recognise and address barriers to equality throughout the process.⁵⁴

Harpur illustrates the difference between these two approaches with reference to an illustrative example of two e-book platforms, where one offered access to the same titles at the same price, but embraced universal design while the other was inaccessible to individuals with disabilities.⁵⁵ Harpur notes that, in terms of the traditional antidiscrimination model, a duty holder could generally purchase either system and consider disability access after the system was in place. In contrast, the management-based approach would incorporate disability accessibility into the decision-making process before purchasing the system.⁵⁶ The explanation by Harpur also accords with the understanding of the positive duties to eliminate discrimination adopted by the court in *R (Bridges) v South Wales Police*. It is contended that this management-based approach to anti-discrimination requires an exercise of due diligence in preventing discrimination from ever occurring in the first place rather than seeking to address the negative effects thereof after it has already occurred.

The Organisation for Economic Co-operation and Development (OECD) has also recently noted that the potential human rights impact of AI requires debates centred around the adoption of AI to be firmly rooted within discussions on businesses' human rights due diligence obligations.⁵⁷ Due diligence is a tailored process that requires a company to address the specific risks within its own operations. The OECD notes that the effect thereof is that within the AI space, the level of due diligence expected of companies would vary widely depending on various factors such as the nature of the AI product being developed and who the users or potential users of that product would be.⁵⁸ Companies developing or deploying AI which has the capacity to be used in harmful ways are required to exercise a greater degree of due diligence to prevent such harms.⁵⁹

The OECD makes a number of important recommendations for AI companies based on its guidance on human rights due diligence.⁶⁰ While the present authors acknowledge the value of such guidelines on responsible corporate conduct, the OECD itself has acknowledged that mandatory laws on due diligence have been more effective.⁶¹ Therefore, these guidelines should ideally be supported by mandatory legislation requiring businesses to conduct appropriate due diligence checks when developing or deploying AI. The implementation of

⁵⁴ Paul Harpur, *Discrimination, Copyright and Equality* (CUP 2019).

⁵⁵ *ibid* 238.

⁵⁶ *ibid.*

⁵⁷ OECD, 'Human Rights Due Diligence Through Responsible AI' in *AI in Business and Finance: OECD Business and Finance Outlook 2021* (OECD 2021).

⁵⁸ *ibid* 74.

⁵⁹ *ibid.*

⁶⁰ *ibid.*

⁶¹ *ibid* 83.

such legislation may indeed be essential for the state to give effect to its own due diligence obligations to regulate the conduct of private parties as required by international human rights law (IHRL).⁶² As previously noted, the ICERD also imposes an explicit obligation on states to eliminate discrimination by private persons.

While laws prohibiting discrimination have become fairly common in most jurisdictions,⁶³ states have been slower to enact provisions requiring businesses to take positive measures to eliminate or prevent discrimination within their organisations other than in specific fields such as employment law. The Promotion of Equality and Prevention of Unfair Discrimination Act (PEPUDA)⁶⁴ in South Africa is somewhat unique in extending this duty to eliminate and prevent racial discrimination to all persons, that is, both natural and legal persons. Section 28(3) of the PEPUDA provides that:

- (a) The State, institutions performing public functions and *all persons have a duty and responsibility*, in particular to:
 - (i) *eliminate discrimination on the grounds of race, gender and disability;*
 - (ii) *promote equality in respect of race, gender and disability.*
- (b) In carrying out the duties and responsibilities referred to in paragraph (a), the State, institutions performing public functions and, where appropriate and relevant, juristic and non-juristic entities, must:
 - (i) audit laws, policies and practices with a view to eliminating all discriminatory aspects thereof;
 - (ii) *enact appropriate laws, develop progressive policies and initiate codes of practice in order to eliminate discrimination on the grounds of race, gender and disability;*
 - (iii) adopt viable action plans for the promotion and achievement of equality in respect of race, gender and disability; and
 - (iv) give priority to the elimination of unfair discrimination and the promotion of equality in respect of race, gender and disability.⁶⁵

The wording of this section is virtually identical to the provision interpreted by the court in *R (Bridges) v South Wales Police*, albeit that the section in that case only

⁶² See eg International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR), art 2(2), on the state's obligation to adopt legislation to give effect to human rights; this duty has also been repeatedly emphasised by regional human rights bodies such as the African Commission on Human and Peoples' Rights (ACCommHPR) in *Zimbabwe Human Rights NGO Forum v Zimbabwe* Communication No 245/2002 (25 May 2006), para 147.

⁶³ See eg the Racial Discrimination Act 1975 in Australia; the Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000 in South Africa; and the Equality Act 2010 in the United Kingdom.

⁶⁴ Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000 (PEPUDA).

⁶⁵ ibid (emphasis added).

imposed obligations on the state. The obligation in PEPUDA clearly requires more from businesses than not discriminating against any person based on race as ‘all persons’ have a positive obligation to promote equality and to eliminate discrimination. This positive duty also requires the adoption of policies and codes aimed at eliminating racial discrimination. When adopting or developing AI products, South African companies could be held to a duty of due diligence in terms of this section, just like the state can be held to such a duty in other jurisdictions such as the United Kingdom (UK) and Australia.⁶⁶ The PEPUDA is also particularly useful because it is not restricted to a specific relationship such as an employment relationship.⁶⁷ Fredman has argued that the traditional focus on defined relationships is a key weakness of discrimination law.⁶⁸ According to Fredman, discrimination law must focus on everyone who impacts inclusion rather than just protecting individuals within a defined relationship such as an employment relationship.⁶⁹

Nevertheless, positive duties are slightly more commonplace in instances where a specifically defined relationship exists, such as in employment law. For example, in the Australian state of Victoria, the Equal Opportunity Act 2010⁷⁰ creates a positive duty for all employers to take reasonable and proportionate measures to eliminate discrimination. This includes discrimination in the screening of applicants for employment. The obligation to take positive measures could reasonably be interpreted to impose a legal duty on an employer who uses AI to screen applicants to conduct a due diligence assessment to ensure that the system will not result in racial discrimination.

Therefore, positive obligations that are only applicable within a defined relationship are not without value. However, the present authors agree with Fredman that addressing widespread racism could require more general obligations that apply beyond the sphere of a specifically defined relationship. This is also important within the context of AI where persons ultimately affected by the use of AI may have no direct relationship with the developer of the AI.⁷¹ In requiring positive measures to address discrimination, the PEPUDA promotes the management approach to addressing discrimination. Therefore, it would not be necessary for a litigant to prove that the AI was discriminatory but merely that there was a potential for discrimination and that insufficient measures are being taken to prevent this risk of discrimination.

⁶⁶ See n 63.

⁶⁷ PEPUDA, s 5(1).

⁶⁸ Sandra Fredman, *Discrimination and Human Rights* (OUP 2004).

⁶⁹ *ibid* 27.

⁷⁰ Equal Opportunity Act 2010 (Vic) (Act No 16 of 2010).

⁷¹ Eg when looking at lethal autonomous weapons systems (LAWS), the customer would likely be the state and not the persons negatively affected by the AI. See Christof Heyns, ‘Human Rights and the Use of Autonomous Weapons Systems (AWS) During Domestic Law Enforcement’ (2016) 38 Human Rights Quarterly 350.

5 Conclusion

This chapter has showcased the growing concern over racial discrimination in the field of AI. The ICERD imposes clear obligations on states to address racial discrimination in all its forms. This obligation is not limited to a negative obligation not to discriminate but also involves a clear positive obligation to eliminate discrimination by its organs and private persons. The chapter draws on the case of *R (Bridges) v South Wales Police* to illustrate the value of positive obligations in addressing the risk of racial discrimination in AI.

This chapter argues that the interpretation of positive obligations in the case largely accords with Harpur's management-based approach to discrimination law. In terms of this approach, persons are required to take active measures to prevent discrimination from occurring rather than only addressing discrimination after it has already adversely affected people. This requires due diligence from both government and private entities to consider possible discriminatory effects of any new AI developed or deployed and the implementation of measures to address these risks.

The authors acknowledge that due diligence does not guarantee that racial discrimination will not occur. However, it is a valuable tool to reduce the risk of developing or deploying discriminatory AI. Due diligence also involves a continuous evaluation process,⁷² meaning that the obligation does not begin and end at the time of developing or deploying the AI system. Should the AI subsequently lead to racially discriminatory results, new measures must be taken to address these risks despite the initial implementation of reasonable measures to prevent discrimination.

However, the chapter has also reflected on the limited number of laws imposing positive obligations on businesses to eliminate racial discrimination. The increase of such laws in specific fields such as employment law is cause for early optimism that states are showing a degree of willingness to impose such obligations on businesses. These states may perhaps, in the future, expand the duties' scope to apply more broadly outside of these defined relationships, such as the generic due diligence duties detectable in the PEPUDA in South Africa.

⁷² OECD (n 57) 67.

Artificial Intelligence and Gender-Based Discrimination

Fabian Lütz

1 Introduction

Algorithms undeniably create a number of challenges for human rights and gender discrimination.¹ The public witnessed women receiving lower credit card limits and discrimination by recruitment algorithms. Internet searches for ‘CEO’ tend to show only pictures of men while searches for ‘nurse’ only those of women.² Such stereotyped and discriminatory outcomes of algorithms either result from the design of artificial intelligence (AI) or, more often, from biased data sets.

The Office of the United Nations High Commissioner for Human Rights (OHCHR) highlighted these potential discriminatory effects and human rights violations of artificial intelligence (AI) systems, notably warning that ‘advances in new technologies must not be used to erode human rights, deepen inequality or exacerbate existing discrimination’.³ The Commissioner underlined that ‘a human rights-based approach to AI requires the application of a number of core principles, including equality and non-discrimination’.⁴ These principles can be found in the Preamble to, and article 1 of, the UN Charter, article 2 of the Universal Declaration of Human Rights (UDHR), the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW), as well as in the human rights framework of the Council of Europe (CoE), the European Union (EU), and national jurisdictions.⁵ Finally, the UN Sustainable Development Goals (SDG), notably SDG5

¹ Mark Latonero, ‘Governing Artificial Intelligence: Upholding Human Rights & Dignity’ (2018) *Data & Society* 1, 10; Eirini Ntoutsi and others, ‘Bias in Data-Driven Artificial Intelligence Systems: An Introductory Survey’ (2020) 10 *WIREs Data Mining and Knowledge Discovery* 1356.

² Fabian Lütz, ‘Gender Equality and Artificial Intelligence in Europe: Addressing Direct and Indirect Impacts of Algorithms on Gender-Based Discrimination’ (2022) 23(1) *ERA Forum* 37.

³ United Nations High Commissioner for Human Rights (OHCHR), ‘The Right to Privacy in the Digital Age’ (2021) UN Doc A/HRC/48/31, para 38.

⁴ *ibid* para 4.

⁵ Council of Europe (CoE), Parliamentary Assembly, Committee on Equality and Non-Discrimination, ‘Report Preventing Discrimination Caused by the Use of Artificial Intelligence’ (2020) Doc 15151, para 66; CoE, Recommendation CM/Rec(2020)1 of the Committee of Ministers to member states on the human rights impacts of algorithmic systems; European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts’ (Artificial Intelligence Act/AI

designed to end all forms of discrimination against all women provide relevant guidance.⁶

Despite the ubiquity of AI⁷ and the UN's ambition for adequate gender equality legislation,⁸ issues of AI and discrimination were only recently raised.⁹ Although a limited number of policy and legal proposals covering gender impacts of algorithms are surfacing, there are currently no specific laws on AI and discrimination, except a local law in New York City that addresses questions of bias and discrimination within automated employment decision tools.¹⁰ However, besides building on existing instruments, such as the CEDAW, arguably specific laws are necessary to enable human rights law (HRL) to satisfactorily address both positive and negative consequences of algorithms.¹¹

In order to evaluate the need for future regulation, this chapter first discusses the negative and positive effects of AI on gender discrimination (section 2). Then, the existing regulatory framework, its shortcomings, as well as the key actors of the UN are discussed. Based on this assessment, possible avenues and elements of a future legal framework are presented, notably, a hybrid approach combining general principles at the UN level and more specific local legal frameworks (section 3).

2 Algorithms' Negative and Positive Effects for Gender Discrimination

While algorithmic gender discrimination might not always be intentional, it mostly results from unrepresentative, incomplete, or biased data sets.¹² Algorithms are not

Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

⁶ United Nations General Assembly (UNGA), UNGA Res 70/1 'Transforming Our World: The 2030 Agenda for Sustainable Development' (25 September 2015) UN Doc A/RES/70/1, goal 5.1.

⁷ Serge Abiteboul and Gilles Dowek, *The Age of Algorithms* (CUP 2020) 1; Panos Louridas, *Algorithms* (MIT Press 2020) 1; Markus Dirk Dubber, Frank Pasquale, and Sunit Das, *The Oxford Handbook of Ethics of AI* (OUP 2020) 253–57.

⁸ UNGA Res 70/1 (n 6) goal 5.c; art 2(b) of the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), New York, 18 December 1979; OHCHR (n 3).

⁹ United Nations Human Rights Council, 'Rights of Persons with Disabilities, Report of the Special Rapporteur on the Rights of Persons with Disabilities' (2022) UN Doc A/HRC/49/52; OHCHR (n 3); T Achiume, 'Racial Discrimination and Emerging Digital Technologies: Report of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance' (2020) UN Doc A/HRC/44/57; UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (2021) UNESCO Doc SHS/BIO/PI/2021/1, paras 87–93.

¹⁰ New York City Int 1894–2020, 'A Local Law to amend the administrative code of the city of New York'; see also California, Bill ACR-125 'Bias and Discrimination in Hiring Reduction Through New Technology' (2019–2020, not adopted).

¹¹ OHCHR (n 3) para 4; Secretary General of the United Nations, *The Highest Aspiration: A Call to Action for Human Rights* (2020) 11 <<https://digitallibrary.un.org/record/3903859>>.

¹² Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law' (2021) 123 West Virginia Law Review 735.

neutral but only as accurate as the underlying (training) data, which reflect existing gender inequalities and stereotypes of society.¹³ Despite the seeming prevalence of negative effects of algorithms for gender equality, they also offer potential positive effects and opportunities from a HRL perspective.¹⁴ Hence, this section will first outline the former aspect (section 2.1) and, subsequently, discusses the latter aspect (section 2.2).

2.1 Negative Effects

Gender biases and discriminations can result from the design of the algorithm,¹⁵ for example, when AI developers model the algorithms based on stereotypical users (eg men).¹⁶ However, biased data sets remain the central concern, notably as proxies can be used to infer gender.

First, AI can facilitate information inference from data sets which 'supports the sorting of individuals into categories, thereby reinforcing different forms of... legal and economic segregation and discrimination'.¹⁷ For example, while the inclusion of the protected characteristic 'gender' might lead to discrimination, omitting or not using 'gender' in data sets and algorithms does not avoid discrimination because 'gender' is substituted by proxies through data correlation.¹⁸ Examples of proxies include the use of specific terms in job applications or CVs, gendered language, salaries, or the practice of redlining which replaces a protected characteristic such as race by postcodes. Rather than using gender expressly, the conscious or unconscious use of those proxies facilitates gender-based discrimination.

Second, in addition, the availability and representativeness of data sets are particularly problematic for gender as illustrated by the *gender data gap*.¹⁹ Due to the digital gender divide, women have less access than men to digital technologies and produce less data to be used by algorithms. This lack of 'female' data leads to

¹³ CoE (n 5) para 3; European Commission, 'Opinion on Artificial Intelligence and Gender Equality of the Advisory Committee on Equal Opportunities for Women and Men' (2020) 4.

¹⁴ CoE planned legal instrument (by 2025) on impacts of AI systems, their potential for promoting equality and risks for non-discrimination, see Council of Europe's Work in progress, Other AI initiatives, including future activities (2023) <www.coe.int/en/web/artificial-intelligence/work-in-progress#02EN>, 4.4 Non-discrimination, Gender equality.

¹⁵ CoE (n 5) para 4; Lorna McGregor, Daragh Murray, and Vivian Ng, 'International Human Rights Law as a Framework for Algorithmic Accountability' (2019) 68 International and Comparative Law Quarterly 309.

¹⁶ UNESCO, *Artificial Intelligence and Gender Equality: Key Findings of UNESCO'S Global Dialogue* (2020) UNESCO Doc GEN/2020/AI/2 REV, 14; Sareeta Amrute, 'Of Techno-Ethics and Techno-Affects' (2019) 123 Feminist Review 56, 70.

¹⁷ CoE, 'Declaration by the Committee of Ministers on the Manipulative Capabilities of Algorithmic Processes' (2019) Decl (13/02/2019)1, para 6.

¹⁸ Ntoutsi (n 1) 4; CoE (n 5) para 3.

¹⁹ Ntoutsi (n 1) 4; Caroline Criado Perez, *Invisible Women: Exposing Data Bias in a World Designed for Men* (Random House 2019); Mayra Buvinic and Ruth Levine, 'Closing the Gender Data Gap' (2016) 13 Significance 34.

unbalanced and non-representative data sets, which paves the way for inaccuracies and inequalities.

The UN ascertains that such ‘biased data sets that lead to discriminatory decisions based on AI systems are particularly concerning’.²⁰

This is exacerbated because algorithms and machine learning (ML) rely on language and images which incorporate gender biases.²¹ This is illustrated by ML techniques such as ‘Word2Vec’ assessing words and language²² using word vectors to show resemblance and associations between words. For example, the word vector ‘Man is to computer programmer as woman is to homemaker?’ reveals gender stereotypes and biases which distort reality, perpetuate stereotypes, and shape future perceptions of gender equality.²³

Problematic for gender discrimination is the algorithms’ lack of accuracy when identifying data correlations. While correlation is understood as ‘a relation existing between phenomena … which tend to … be associated, or occur together in a way not expected on the basis of chance alone’, causation is ‘the act or process of causing’.²⁴ As a matter of illustration, recruitment algorithms could find patterns and correlations but not necessarily causal links between information provided in application dossiers or CVs. Due to the *gender data gap*, the scrutinised variables may lead to gender discrimination, as evidenced by the private recruitment algorithms or public administrations’ algorithms for unemployment-related benefits.²⁵

In order to mitigate these effects of the gender data gap and proxy discrimination, regulation should therefore impose testing and evaluation for biases and (gender) discrimination for algorithms and training data sets.

²⁰ UN (n 3) para 19.

²¹ See Ryan Steed and Aylin Caliskan, ‘Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases’ (2021) Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 701.

²² Ethem Alpaydin, *Machine Learning* (MIT Press 2021) 133–35; Stuart Russel and Peter Norvig, *Artificial Intelligence: A Modern Approach* (4th edn, Pearson 2021) 908, 926–29.

²³ Tolga Bolukbasi and others, ‘Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings’ (2016) arXiv:1607.06520, 1–3.

²⁴ ‘Correlation,’ ‘Causation,’ entries in *Merriam-Webster dictionary*, <www.merriam-webster.com/dictionary>; Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books 2018) 5.

²⁵ In general, see Lütz (n 2) 39; CoE (n 5) para 17; Jeffrey Dastin, ‘Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women’ (*Reuters*, 11 October 2018) <www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>; Doris Allhutter and others, ‘Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics are Made Effective’ (2020) 3(5) *Frontiers in Big Data* 1–2; Adamantia Rachovitsa and Niclas Johann, ‘The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case’ (2022) 22 *Human Rights Law Review* 1.

2.2 Positive Effects

Despite machine biases and the assumption that humans need to remain in control to safeguard human rights,²⁶ algorithms still have several advantages—such as coherence, objectivity, and speed—which enable them to contribute to achieve gender equality and better enforcement of non-discrimination law.²⁷ One example is the automatic detection of biases and gender discrimination with the help of algorithms, which has been used for detecting gender biases in YouTube videos.²⁸ Benefits could also arise where regulators use algorithms in their enforcement efforts by relying on the strengths of technology where humans face decision-making bias and noise.²⁹ Even though algorithms are not unbiased, they are helpful to overcome or reduce (some of the) gender biases and noise.³⁰ They could be used to assist companies and administrations to make gender biases visible and to ensure coherence, reliability, equal outcomes, and thereby help to ensure better and less biased private and public decisions. To this end, computer scientists should work jointly with non-discrimination experts to better understand the effects of and the usage of AI. Algorithms could also help enforcers to protect human rights by detecting discrimination, for example, by testing other algorithms for potential discrimination,³¹ or by automatically screening relevant case law to facilitate complaint handling. Such positive effects could be generated by complementing assessment and decisions of public officials to achieve better enforcement of discrimination rules both in the offline and online world.

3 Existing and Future Avenues to Mitigate Negative Effects of Algorithms

Despite the existence of a general legal framework on gender discrimination, this chapter will argue that AI's positive and negative impacts on gender equality could be better addressed in specific legal instruments.³² While soft law instruments have an undeniable merit, their many limitations to ensure the principle of non-discrimination advocate for a legal human rights-based framework that addresses

²⁶ See Hannah Fry, *Hello World: How to be Human in the Age of the Machine* (Random House 2018) 76–81; Frank Pasquale, *New Laws of Robotics: Defending Human Expertise in the Age of AI* (Belknap Press 2020); AI Act, art 14.

²⁷ UNGA Res 70/1 (n 6) goal 5.b.

²⁸ CoE (n 5) para 36.

²⁹ Kai-Fu Lee, *AI 2041: Ten Visions for the Future* (Currency 2021) xiv; UN (n 3), para 58; Daniel Kahneman, ‘Noise’ (2016) Harvard Business Review 38, 43.

³⁰ Daniel Kahneman, Olivier Sibony, and Cass R Sunstein, *Noise: A Flaw in Human Judgment* (Little Brown 2021) 334–37.

³¹ UN (n 9).

³² CoE (n 5) para 14.

the impacts of AI on gender discrimination by adapting appropriate legal instruments at the UN and national levels.

This section will first look at the existing regulatory framework and observe the lack of specific hard law to regulate algorithmic discrimination. Subsequently, after having discussed the limitations of soft law and existing proposals, possible avenues to address the shortcomings and relevant UN actors will be presented.

3.1 The Existing Framework and Its Limitations

The international regulatory framework provides for the protection of gender-based discrimination but there are no specific legal instruments yet that address algorithmic gender discrimination aside from political declarations by UN human rights bodies.³³ While existing hard and soft law have limitations when it comes to capturing the realities of algorithmic discrimination, the UN actors are well placed to address arising issues of AI gender discrimination if equipped with the right tools.

3.1.1 Hard Law

Both UN and local frameworks generally provide for the protection against gender discrimination.

At the UN level, the CEDAW is a comprehensive global human rights framework for the protection against gender discrimination. The CEDAW is more than four decades old, but it could be an anchor to cover algorithmic discrimination, either via a broad and generous interpretation by the CEDAW Committee or alternatively, by reforming this UN Convention to specifically address the AI challenges for gender equality.³⁴ Article 2 of the CEDAW provides a legal framework that could be used to address new forms of gender-based discrimination. The word ‘all’ in article 2 of the CEDAW—‘States Parties condemn discrimination against women in *all* its forms, agree to pursue by all appropriate means and without delay a policy of eliminating discrimination against women’—could be interpreted as covering algorithmic discrimination. On this basis the UN could invite (signatory) states to adopt legislation specifically covering AI.

With regard to some negative effects explained in section 2, such as biased outcomes of algorithms, article 5(a) of the CEDAW could help address biases, prejudices, and stereotyped roles for men and women by calling upon states ‘to modify the social and cultural patterns of conduct of men and women, with a view

³³ Christiaan van Veen and Corinne Cath, ‘Artificial Intelligence: What’s Human Rights Got To Do With It?’ (2018) 14 Data & Society.

³⁴ CoE (n 5) para 63.

to achieving the elimination of prejudices and customary and all other practices which are based ... on stereotyped roles for men and women.³⁵

Regarding the CoE framework, article 14 of the European Convention on Human Rights (ECHR) prohibits gender discrimination albeit only in relation to the enjoyment of the rights and freedoms of the Convention. Article 1(1) of Protocol No 12 to the ECHR establishes a self-standing prohibition of discrimination: ‘The enjoyment of any right set forth by law shall be secured without discrimination on any ground such as sex’.³⁶ As the ECHR is a ‘living instrument’,³⁷ one could imagine future jurisprudence covering AI discrimination but to date, relevant case law remains rare in relation to algorithms.³⁸

Similarly, many regional frameworks, for example, in Australia, the EU, the United States (US), the Organization of American States (OAS), the African Union (AU), or the Association of Southeast Asian Nations (ASEAN) address, in principle, gender-based discrimination but have not addressed specifically algorithmic discrimination.³⁹

One of the rare examples of local legislation that will address gender biases and discrimination, albeit limited to AI recruitment tools, entered into force in January 2023 in the City of New York. In essence, it mandates not only bias audits before companies and public authorities can use AI recruitment tools, but also foresees information requirements, transparency rules, and penalties for violations of the law.⁴⁰

3.1.2 Soft Law

In terms of soft law, some states have adopted soft law frameworks on AI, but they are neither binding nor do they specifically address gender discrimination.⁴¹ At the international level, the UN Guiding Principles on Business and Human Rights⁴²

³⁵ Lisa Baldez, ‘The UN Convention to Eliminate All Forms of Discrimination Against Women (CEDAW): A New Way to Measure Women’s Interests’ (2011) 7 Politics & Gender 419, 423.

³⁶ Protocol No 12 to the Convention for the Protection of Human Rights and Fundamental Freedoms, ETS No 177.

³⁷ ECHR, ‘European Convention on Human Rights: Living Instrument at 70’ <https://echr.coe.int/Documents/Convention_Instrument_ENG.pdf>.

³⁸ *Big Brother Watch v the United Kingdom* App Nos 58170/13, 62322/14, and 24960/15 (ECtHR, 25 May 2021), para 159.

³⁹ See Australia: Sex Discrimination Act 1984, No 4; Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) [2006] OJ L204/23–36; Title VII of the Civil Rights Act of 1964, Pub L 88-352 (Title VII); Inter-American Convention Against All Forms of Discrimination And Intolerance (A-69) (adopted on 5 June 2013); article 2 of the African Charter on Human and Peoples’ Rights; Protocol to the African Charter on Human and People’s Rights on the Rights of Women in Africa; and the ASEAN Human Rights Declaration.

⁴⁰ See note 10, paras 20-870, 20-871 of the New York City law.

⁴¹ Canada: Directive on Automated Decision-Making (2021); (US) National AI Initiative Act (2020), Division E, Sec 5001 (suggests ‘methods to assess, characterize, and reduce bias in datasets and artificial intelligence systems’, 39); Australia: Artificial Intelligence Ethics Framework and Artificial Intelligence Action Plan (2021).

⁴² Human Rights Council, ‘Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John

(UNGPs) which aim to prevent, address, and remedy human rights abuses committed in business operations, are a good example where principles are addressed to both states and companies. The UNGPs require states to adopt laws that ensure the respect of human rights by businesses and periodically assess adequacy and gaps, specifically mentioning non-discrimination laws and guidance, including human rights due diligence and issues of gender.⁴³ Although consensual adoption by the Human Rights Council (HRC) grants them high legitimacy, the UNGPs have a low enforceability as they depend on states to implement rules in national law. Nevertheless, the UNGPs and the UN reports that call on states to act and encourage businesses to address human rights abuses can facilitate and encourage the advancement of gender equality, also in the algorithmic age. The UN Special Rapporteur, for example, recommended to combat discrimination linked to AI.⁴⁴

Specifically on AI, the CoE issued a recommendation on the human rights impacts of algorithmic systems and the OECD adopted a recommendation on AI.⁴⁵ Both contain several elements regarding non-discrimination and impose obligations on states and private actors. For example, businesses ‘should … ensure that the design, development, and ongoing deployment of their algorithmic systems do not have direct or indirect discriminatory effects’.⁴⁶ In addition, both soft law instruments emphasise that data sets should be non-discriminatory, and testing and evaluation are key for avoiding unjustified discriminatory impacts.⁴⁷

Many companies or professional organisations also have self-binding principles or AI standards with specific aims on discrimination and gender equality.⁴⁸ These standards coexist with more independent and global standards established by the International Organization for Standardization (ISO).⁴⁹ Despite concretely addressing issues of bias and discrimination and being global in their application,

Ruggie—Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework’ (21 March 2011) UN Doc A/HRC/17/31; United Nations Development Programme, *Gender Dimensions of the Guiding Principles on Business and Human Rights* <<https://www.undp.org/publications/gender-dimensions-guiding-principles-business-and-human-rights>>.

⁴³ HRC, ‘Guiding Principles on Business and Human Rights’ (16 June 2011) UN Doc A/HRC/17/31, 8.

⁴⁴ ibid para 61.

⁴⁵ CoE (n 5); OECD, *Recommendation of the Council on Artificial Intelligence* (2019) OECD/LEGAL/0449; Karen Yeung, ‘Recommendation of the Council on Artificial Intelligence (OECD)’ (2020) 59 International Legal Materials 27, 32.

⁴⁶ CoE (n 5) para C1.4; OECD (n 45) 1.2(a).

⁴⁷ CoE (n 5) paras 2.2, 3.3, 3.4; OECD (n 45) 1.4.

⁴⁸ Google, *AI Principles* 2020 <<https://ai.google/responsibility/principles/>>; Facebook’s Five Pillars of Responsible AI 2019 <<https://ai.facebook.com/blog/facebook-s-five-pillars-of-responsible-ai/>>; Microsoft, *Guidelines for Human-AI Interaction* 2019 <www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/>.

⁴⁹ Jessica Fjeld and others, ‘Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI’ (Berkman Klein Center for Internet and Society, 2020); ISO Standards <www.iso.org/committee/6794475.html>.

such codes of conduct, best practices, and standards are not legally binding and deprive victims of gender discrimination of legal remedies.⁵⁰

It is questionable whether current frameworks are sufficient to uphold human rights. The existing hard law could address some of the problems of algorithmic discrimination, notably if judges and relevant administrative bodies will interpret the existing gender equality rules. However, to ensure legal certainty and effective enforcement of gender equality law and take into account the specificities of AI, it seems more adequate to adopt specific rules that address the negative effects associated with AI (described in section 2), rather than waiting for jurisprudence to clarify issues on a case-by-case basis.

These soft law instruments, standards, and guidelines could cover the impact of AI on gender discrimination and would apply to a large number of firms in many countries and consequently could protect more people than national laws. Furthermore, industry knowledge and a good understanding of AI systems could lead to standards reflecting the issues of bias and discrimination. A major limitation of soft law instruments is non-enforceability in courts or appropriate follow up in HRL mechanisms necessary to ensure the respect of gender equality.⁵¹

Consequently, despite their advantages, soft law measures have clear limitations and cannot substitute a legal framework addressing the risks of gender discrimination associated with AI. The doctrine mainly agrees with the shortcomings of existing hard and soft law frameworks and sees added value of new regulation to protect human rights⁵² but some authors argue against the creation of AI specific non-discrimination laws.⁵³ However, the UN could rely on soft law, the technical expertise contained in company guidelines, or ISO standards when building legal frameworks and make use of the existing institutional actors involved in gender equality policy.

⁵⁰ CNCDH, *Avis relatif à l'impact de l'intelligence artificielle sur les droits fondamentaux* (2022) A-2022-6, paras 51–52.

⁵¹ Justine Nolan, ‘The Corporate Responsibility to Respect Rights: Soft Law or Not Law?’ in Surya Deva and David Bilchitz (eds), *Human Rights Obligations of Business: Beyond the Corporate Responsibility to Respect* (CUP 2013) 138; Daniel Augenstein, ‘Negotiating the Hard/Soft Law Divide in Business and Human Rights: The Implementation of the UNGPs in the European Union’ 9 Global Policy 254.

⁵² Frederik J Zuiderveen Borgesius, ‘Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence’ (2020) 24(10) International Journal of Human Rights 1572–93, 1586; Philip Hacker, ‘Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law’ (2018) 55(4) Common Market Law Review 1143, 1161–70.

⁵³ Janneke Gerards and Frederik Zuiderveen Borgesius, ‘Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence’ (2022) 20 Colorado Technology Law Journal 1.

3.1.3 UN Actors

A prominent role regarding the protection against gender-based algorithmic discrimination should be played by existing UN actors,⁵⁴ but also by regions⁵⁵ and states.

The United Nations General Assembly (UNGA)⁵⁶ has a clear human rights mandate which could deliver clear political statements or resolutions to frame the discussion on AI and gender discrimination. The OHCHR, for example,⁵⁷ is ‘mandated by the UNGA to promote and protect the enjoyment and full realization, by all people, of all human rights’⁵⁸ and could help increase protection against gender discrimination and ‘engage in a dialogue with all governments in the implementation of his/her mandate with a view to securing respect for all human rights’.⁵⁹ Subsidiary human rights organs, such as the HRC,⁶⁰ address questions of gender equality frequently including some issues of AI and its impacts on gender.⁶¹ The Commission on the Status of Women (CSW)⁶² could serve as a laboratory for ideas on how to ensure human rights protection when AI impacts gender equality, for example, during its 67th CSW session covering, *inter alia*, innovation, technological change, and achieving gender equality. CSW is the principal global inter-governmental body exclusively dedicated to promoting gender equality and the empowerment of women, which adopts recommendations in the form of negotiated and agreed conclusions during its yearly sessions. Finally, UN organs monitoring treaty compliance, such as the Human Rights Committee or the Committee on the Elimination of Discrimination Against Women⁶³ could help to avoid algorithmic discrimination and monitor compliance with any future treaty on discrimination and AI. The CEDAW could play a vital role via its monitoring and enforcement mechanisms⁶⁴ and its interpretation and application of the right to non-discrimination and equality.⁶⁵ The Optional Protocol to the CEDAW which created the communications procedure (article 2) and the inquiry procedure

⁵⁴ Frédéric Mégré and Philip Alston, *The United Nations and Human Rights: A Critical Appraisal* (OUP 2020).

⁵⁵ See Inter-American Commission of Women (CIM) for Latin America or the Women, Gender, Development and Youth Directorate (WGDY) for the African Union (AU), both responsible for promoting gender equality and addressing gender discrimination.

⁵⁶ Mégré and Alston (n 54) 99–131.

⁵⁷ *ibid* 667–709.

⁵⁸ UNGA, Res 48/141 (20 December 1993) UN Doc A/RES/48/141.

⁵⁹ *ibid* point 4(g).

⁶⁰ Mégré and Alston (n 54) 181–239.

⁶¹ HRC (n 9).

⁶² Mégré and Alston (n 54) 253–91; UN Economic and Social Council Resolution 11(II) (21 June 1946) UN Doc E/RES/11(II).

⁶³ Mégré and Alston (n 54) 393–439.

⁶⁴ Bal Sokhi-Bulley, ‘The Optional Protocol to CEDAW: First Steps’ (2006) 6 *Human Rights Law Review* 143, 144–45.

⁶⁵ Simone Cusack and Lisa Pusey, ‘CEDAW and the Rights to Non-Discrimination and Equality’ (2013) 14 *Melbourne Journal of International Law* 54; Neil A Englehart and Melissa K Miller, ‘The CEDAW Effect: International Law’s Impact on Women’s Rights’ (2014) 13 *Journal of Human Rights* 22.

(article 8) grant new enforcement possibilities to the CEDAW Committee to explore algorithmic discrimination.⁶⁶

Thus, UN actors could shape the political discourse and develop legal rules that address the positive and negative impacts of AI on gender discrimination. Within a hybrid approach, those general principles could be adopted at the regional or national levels and further refined in order to ensure effective enforcement.

Section 3.2 outlines the role of international HRL to address algorithmic discrimination and sketch out elements of a framework on AI and gender discrimination.

3.2 A Hybrid Approach to Address Algorithmic Discrimination

Some authors argue that existing international HRL (as presented in section 3.1) already provides for some elements to cover algorithmic discrimination.⁶⁷ However, as will be argued in section 3.2.1, the nature of AI suggests the need for a new global and hybrid approach to address algorithmic discrimination, combining the general UN legal framework with regional and national legal frameworks for specific rules. Some key elements for a future regulatory framework will be presented.

3.2.1 HRL Framework for Algorithmic Discrimination

International and regional human rights frameworks seem best placed to provide general guidance and principles for algorithmic discrimination.⁶⁸ Leaving implementation at the national level enables the specific design and enforcement respecting local legal traditions.

Opting for a wide regional or global regulation is embedded in the transnational effects of algorithms, because algorithms are designed, programmed, sold, and used in different countries, producing results and legal effects that could affect people beyond national jurisdictions. Different applicable legal regimes create legal uncertainty as to whether algorithmic discrimination is covered. In addition, algorithms use similar training data and are fed with results and data used in other countries to further develop and improve the algorithm. Whereas some countries use classical non-discrimination laws, others might adopt specific regulatory regimes for AI systems that also cover gender discrimination. In other words, while algorithms learn transnationally, regulation remains foremost national or regional.⁶⁹ If algorithms are used by thousands of companies worldwide, it becomes

⁶⁶ Optional Protocol to the Convention on the Elimination of All Forms of Discrimination Against Women, UNGA Res 54/4 (6 October 1999) UN Doc A/RES/54/4.

⁶⁷ McGregor, Murray, and Ng (n 15) 329; Paul Nemitz, 'Constitutional Democracy and Technology in the Age of Artificial Intelligence' (2018) 376 *Philosophical Transactions of the Royal Society* 11.

⁶⁸ For multilateralism, see HA Kissinger and others, *The Age of AI: And Our Human Future* (Little Brown 2021); CoE (n 5) para 4.

⁶⁹ CoE (n 5) para 80.

evident that gender discrimination needs to be adequately addressed at the global level and specified and complemented at the regional or local level.

Using effects of *cross-pollination* at different governance levels (UN, CoE, EU) could facilitate the creation of legal frameworks and a global level-playing field based on human rights.⁷⁰ Regulatory fora—such as the EU, ASEAN, or the AU—tend to act as standard setters for their member states and shape HRL regarding gender equality and AI.⁷¹ Consequently, any future regulatory attempts can build on the existing frameworks. As human rights obligations in treaties bind every signatory state,⁷² references to human rights in AI regulation could give effect to a human rights-based approach on gender discrimination followed globally and implemented in national legal orders.

While UN frameworks relying on general principles could address more countries and companies, states tend to formulate binding rules. As soft law measures often depend on the state's willingness to adopt or ratify legal acts, they guarantee less effective enforcement than national jurisdictions.⁷³ Considering the global reach and transnational nature of AI applications, the added value of combining general principles at the UN level with binding rules at the regional or state level could increase the number of jurisdictions covering AI discrimination. International norms governing AI should therefore coexist with and reinforce regional and national norms to ensure an adequate level of protection against gender discrimination in the algorithmic age.

3.2.2 Elements for Regulation

Many elements of a future regulatory framework aimed at addressing algorithmic discrimination can be drawn from existing proposals, such as the EU's AI Act or the CoE AI recommendation. Both could inspire global AI regulation and serve as blueprint for a non-discrimination-specific instrument at the UN level. Preventing biases and algorithmic discrimination can be achieved in several ways. Different approaches exist, such as prohibition, ex ante, ex post, or risk-based regulation.⁷⁴ In order to ensure compliance with human rights, legislation could also target the

⁷⁰ Thomas Giegerich, *The European Union as Protector and Promoter of Equality* (Springer 2020) 1; Treaty on the Functioning of the European Union (TFEU), art 220(1); OECD (n 45) point VIII and 2.5.

⁷¹ Hao Duy Phan, 'The Evolution Towards an ASEAN Human Rights Body' (2008) 9 Asia-Pacific Journal on Human Rights and the Law 1–12; Gerard Clarke, 'The Evolving ASEAN Human Rights System: The ASEAN Human Rights Declaration of 2012' (2012) 11 Journal of Human Rights 1–27; Arthur Gwagwa and others, 'Artificial Intelligence (AI) Deployments in Africa: Benefits, Challenges and Policy Dimensions' (2020) 26 African Journal of Information and Communication 1, 26.

⁷² James Crawford and Ian Brownlie, *Brownlie's Principles of Public International Law* (OUP 2019) 610.

⁷³ OECD (n 70).

⁷⁴ UN (n 3) para 45; United Nations (UN), 'Urgent Action Needed over Artificial Intelligence Risks to Human Rights' (*UN News*, 15 September 2021) <<https://news.un.org/en/story/2021/09/1099972>>; AI Act, art 6; high-risk AI systems (eg AI-recruitment systems) require specific regulation, see AI Act, arts 6(1), (2), 7(1)(b), Recital 36, and Annex III; see also precautionary principle: TFEU, art 191(2).

AI design stage by prohibiting algorithms unless they fulfil very strict conditions, notably with prior bias audits⁷⁵ or AI impact assessments.⁷⁶ A UN framework should be built on the idea that every hard law can only be effective if adequate sanctions for non-compliance are foreseen.⁷⁷ Any future UN and the planned CoE legal framework (2025)⁷⁸ should build on the existing soft law frameworks.

Algorithmic gender discrimination in the labour market will serve as an example to illustrate some of the elements for a human rights-based framework. Algorithmic discrimination can occur throughout the employment life cycle⁷⁹ and might remain unnoticed, as a potential applicant might not receive information on (pre-)selection.⁸⁰ Whereas humans (can) give reasons and justify rejections for jobs or dismissals, there is no such possibility for algorithms. This complicates evidence gathering and proof for discrimination claims.⁸¹ Opaque AI processes, limited influence by humans, lack of transparency, and data sets used for AI recruitment,⁸² make it more difficult to access relevant evidence of potential discrimination.⁸³ Therefore, transparency and accountability regarding the algorithm and the data sets should be core elements of future regulation, not least for enabling victims to pursue discrimination claims.

The ECHR, for example, imposes a high threshold for evidence and proof of discrimination.⁸⁴ In gender discrimination cases, only 'the disproportionately prejudicial effect on a particular group'⁸⁵ is required, which could facilitate claims

⁷⁵ Ari Ezra Waldman, 'Power, Process, and Automated Decision-Making' (2019) 88 Fordham Law Review 613, 632; B Wagner and others, 'Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications' DGI (2017)12 (prepared by the Committee of Experts on Internet Intermediaries (MSI-NET) for the Council of Europe) 45.

⁷⁶ Alessandro Mantelero, 'AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment' (2018) 34 Computer Law & Security Review 754, 757; Brandie Nonnecke and Philip Dawson, 'Human Rights Implications of Algorithmic Impact Assessments: Priority Considerations to Guide Effective Development and Use' (Carr Center for Human Rights Policy Harvard Kennedy School, 2021); UN (n 4) para 60(a); McGregor, Murray, and Ng (n 15).

⁷⁷ See AI Act, art 71 and para 20-872 of New York's law Int 1894-2020.

⁷⁸ CoE (n 5).

⁷⁹ UNESCO (n 9) para 87; Josephine Yam and Joshua August Skorburg, 'From Human Resources to Human Rights: Impact Assessments for Hiring Algorithms' (2021) 23 Ethics and Information Technology 611.

⁸⁰ NYC mandates prior disclosure and bias audits for AI-recruitment (n 10); GDPR, art 22(1), (2) (c), (3).

⁸¹ Samantha Besson, 'Evolutions in Non-Discrimination Law within the ECHR and the ESC Systems: It Takes Two to Tango in the Council of Europe' (2021) 60 American Journal of Comparative Law 147, 168.

⁸² UN (n 3) paras 33, 60; CoE (n 5) paras 4.1, 4.3; AI Act, arts 13, 52.

⁸³ See Case C-104/10 *Kelly* [2011] ECR I-6813; Case C-415/10 *Meister* ECLI:EU:C:2012:217; Case C-274/18 *Schuch-Ghannadan* ECLI:EU:C:2019:828, para 57.

⁸⁴ *DH v the Czech Republic* App no 57325/00 (ECtHR, 13 November 2007), para 79; *Velikova v Bulgaria* App no 41488/98 (ECtHR, 8 May 2000), para 94; *Hugh Jordan v the United Kingdom* App no 24746/94 (ECtHR, 4 April 2001), para 154; *Hoogendijk v the Netherlands* App no 58641/00 (ECtHR, 6 January 2005), para 207; *Opuz v Turkey* App no 33401/02 (ECtHR, 9 June 2009), para 198.

⁸⁵ *Hoogendijk* (n 84) para 207; *DH* (n 84) para 188; *Opuz* (n 84); *Orsus v Croatia* App no 15766/03 (ECtHR, 16 March 2010), paras 152–53, 155.

for algorithmic discrimination. Generally, access to evidence and the burden of proof refrain potential claimants from bringing cases because information is in the company's domain.⁸⁶ Without facilitating access to evidence or shifting the burden of proof⁸⁷ non-discrimination claims remain difficult. In addition, as decisions move from humans to algorithms, accountability is often unclear and appropriate legal remedies must be available to victims to launch complaints or court cases.⁸⁸

One way to avoid algorithmic discrimination from falling through the cracks is mandatory assessment and monitoring of algorithms notably through (bias) audits to ensure AI systems are less prone to discriminatory outcomes.⁸⁹

The discussed proposals and opinions on human rights and AI⁹⁰ contain important elements to inspire a UN approach to algorithmic discrimination. While general principles regarding the impacts of AI on gender could be spelled out at the UN level, the more specific implementation could be left for the regional or local level. The UN human rights system could draw on the resources, ideas, and experiences of its actors and the global community to advance concrete proposals covering AI and gender discrimination. This could mitigate the harm potentially caused by AI and exploit the potential of AI for positive effects on GE.

Besides regulation, some of the risks of AI for gender equality could be overcome by political measures, such as information campaigns, promoting diversity, female labour market participation in the STEM/AI sector, and by ensuring awareness among AI programmers of the principle of non-discrimination.⁹¹

4 Conclusion

Algorithms are not neutral or discriminatory per se, it all depends on design and data. In light of the absence or limitations of current regulatory frameworks, specific regulation on AI and discrimination is the solution to safeguard human rights, ensure non-discrimination and gender equality. In general, more collaboration between regional and international actors is also needed. UN instruments must develop general principles that should be mirrored at the national and regional levels. Beyond regulation, algorithms should be used by enforcers to detect and enforce gender discrimination. The increasing awareness of global leaders and recent UN

⁸⁶ Schuch-Ghannadan (n 83).

⁸⁷ Case C-109/88 *Danfoss* (1989) ECR 3199, para 16.

⁸⁸ UN (n 3) para 58–59(g); CoE (n 5) paras 4, 4.5, 6.4; AI Act, arts 17, 71.

⁸⁹ UN (n 4) para 59.

⁹⁰ CNCDH (n 50).

⁹¹ National AI Initiative Act (2020) (n 41) s 2, para 8; CoE (n 5) paras 56–57.

reports creates hope that concrete legal frameworks will follow the political statements. A hybrid approach based on the UN framework, for example, by refining and modernising the CEDAW or creating specific legislative instruments on algorithmic discrimination could be an option. Regional and national actors could then update their respective legislative framework to ensure human rights protection in the area of algorithmic discrimination.

Artificial Intelligence and LGBTQ+ Rights

Masuma Shahid

1 Introduction

The last two decades have been incremental for the development of artificial intelligence (AI) and data-driven technologies. AI is increasingly being applied to all facets of life. Through its vast application and scope, it inevitably has its effects on human rights; these can be both positive as well as negative. Much research has been conducted on AI and its interplay with human rights in general;¹ yet, little research is available on AI's impact on LGBTQ+² rights and the *queer* community in particular.³ While the LGBTQ+ community has benefitted from certain forms of AI, there are also discernible risks involved in terms of privacy, health, safety, employment, censorship, discrimination, among other issues. Weighed against each other, the pertinent question is whether AI is to be considered a friend or a foe for the LGBTQ+ community. This chapter aims to answer that question and fill the gap in the literature. Section 2 discusses the ways in which AI affects human rights. Section 2.1 ventures into the positive examples of AI applications that have improved the lives of LGBTQ+ individuals and advanced or promoted the protection of their rights, while section 2.2 discusses some negative examples and indicates the risks and ethical implications of this ever-evolving technology for LGBTQ+-related rights. Section 3 discusses how sexual orientation and gender

¹ See eg R Walters and M Novak, *Cyber Security, Artificial Intelligence, Data Protection and the Law* (Springer 2021); Alan Turing Institute, 'Artificial Intelligence, Human Rights, Democracy and the Rule of Law: A Primer' (Council of Europe and the Alan Turing Institute 2021); E Aizenberg and J van den Hoven, 'Designing for Human Rights in AI' (2020) 7(2) Big Data & Society 1; J Gerards, 'The Fundamental Rights Challenges of Algorithms' (2019) 37(3) Netherlands Quarterly of Human Rights 205; F Raso and others, *Artificial Intelligence and Human Rights: Opportunities and Risks* (Berkman Klein Center for Internet and Society at Harvard University 2018); S Barocas and AD Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671.

² The term 'LGBTQ+' is used as an overarching yet interchangeable term for the 'queer' community, consisting of people with various kinds of gender, sexual, and romantic identities and attractions. This abbreviation is by no means exhaustive, hence the use of '+'.

³ Notable exceptions are N Tomasev and others, 'Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities' (Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society 2021) 254; and PHC Wilkinson, 'The Legal Implications of Sexual Orientation-Detecting Facial Recognition Technology' (2021) 20 Dukeminier Awards Journal 301, though the latter specifically focuses on facial recognition technology. See also the 'Queer in AI' conferences organised by queer scientists with a view of making the AI community a safe and inclusive place that welcomes, supports, and values LGBTQ+ people, <sites.google.com/view/queer-in-ai/home>.

identity (SOGI)⁴ are protected by the current international and regional regulatory human rights framework when being affected by AI applications. Section 4 discusses what *queer theory* is and how its application from the beginning stages of AI-development on could take away some of the human rights concerns that are prevalent, while the three case studies discussed in section 5 illustrate how an actual *queerification* of AI could lead it to being more inclusive, diverse, and fair for the *queer* community.

2 AI's Impact on Human Rights

The use and application of AI has changed society forever. On the one hand, breakthroughs in AI have improved healthcare, education, transportation, public administration, manufacturing, customer service, and so on. On the other hand, the use of AI can adversely affect human rights. The ways in which this takes place are diverse. In the private sector, more and more insurance companies are using algorithms to pre-screen and assess claims, banks are using software and machine learning (ML) to develop pricing models and assess mortgages for different types of individuals and companies are using software to hire potential candidates. In the public sector, governmental organisations are increasingly using ML to predict the most effective outcomes for certain policies and developing data models. This does not always go well. In their groundbreaking research, Barocas and Selbst describe five ways in which AI decision-making can lead to discrimination;⁵ for instance, in how the data is collected, labelled, distinguished, and so on. In all of these decisions, human biases can get programmed in. Notable examples are the Dutch childcare benefits ('*Toeslagenaffaire*')⁶ and welfare benefits surveillance cases⁷ where the Dutch government was reprimanded for using automated risk assessment tools for flagging potential tax and/or benefits fraud. Thousands of Dutch citizens were falsely flagged based on traits linked to discriminatory biases (eg ethnicity or the possession of dual nationality).⁸ Similarly, Raso and others distinguish five dangerous implications or risks of AI for human rights,⁹ *inter alia*, the uneven distribution of the positive and negative aspects of AI across society and the perpetuation, amplification, and ossification of existing social biases and prejudices, with attendant consequences for the right to equality.¹⁰

⁴ The term 'SOGI' is most commonly used in the United Nations (UN) framework.

⁵ Barocas and Selbst (n 1) 671; see also F Zuiderveen Borgesius, *Discrimination, Artificial Intelligence and Algorithmic Decision-Making* (Directorate General of Democracy of the Council of Europe 2018).

⁶ Tweede Kamer der Staten-Generaal, 'Ongekend Onrecht' ['Unprecedented Injustice'] (Report, 17 December 2020).

⁷ *SyRI case*, Case C-09-550982-HA ZA 18-388 (Court of The Hague, decision of 5 February 2020).

⁸ PricewaterhouseCoopers, 'Onderzoek Effecten FSV Toeslagen' ['Research Into the Effects of Fraud Detection of Benefits'] (Report, November 2021).

⁹ Raso and others (n 1).

¹⁰ *ibid* 17–19.

In sum, these examples demonstrate the need for binding legal framework and enforceable regulation, both at national as well as at international or regional level. Furthermore, in the case of minority groups such as the LGBTQ+ community, special attention needs to be paid to the specific qualities of the groups in order to prevent the compounding effect of intersectional discrimination.¹¹ Section 5 of this chapter focuses on queer theory as a way to ensure that the interests and values of the LGBTQ+ community are considered from the development stages of AI on with the aim of this resulting in more (LGBTQ+) fairness in and by AI.

2.1 Positive Examples of AI Applications Relevant to the LGBTQ+ community

Certain AI applications, especially those in health sciences, have been extremely beneficial for the LGBTQ+ community. One notable example is the partnership between the Trevor Project, Google, and PricewaterhouseCoopers. As suicide rates are generally three times higher in LGBTQ+ youth,¹² the Trevor Project has received millions in grants for AI-powered suicide prevention and risk assessor tools.¹³ Another AI application is the development of ML models by IBM for the promotion of LGBTQ+ mental well-being in the workplace.¹⁴ In addition, ML models and algorithms have been deployed to identify candidates for HIV pre-exposure treatments¹⁵ and outcomes.¹⁶

AI applications have likewise been helpful in the advancement of the protection of LGBTQ+ rights and individuals in different ways. An example is the use of deepfake technology to swap faces and anonymise individuals in a documentary on the LGBTQ+ community in Russia.¹⁷ This technology is innovative and can

¹¹ K Gaunt, ‘Tips for Applying an Intersectional Framework to AI Development’ (*TechCrunch*, 18 December 2020) <<https://techcrunch.com/2020/12/18/tips-for-applying-an-intersectional-framework-to-ai-development/>>.

¹² Centers for Disease Control and Prevention (CDC), ‘Sexual Identity, Sex of Sexual Contacts, and Health-Risk Behaviors among Students in Grades 9–12: Youth Risk Behavior Surveillance’ (US Department of Health and Human Services 2016).

¹³ The Trevor Project, ‘Innovation at the Trevor Project: Using Machine Learning to Assess LGBTQ Youth Suicide Risk with Google.org’ <www.thetrevorproject.org/innovation-the-trevor-project/>; The Business Wire, ‘The PwC Charitable Foundation Inc Invests \$6 million in the Trevor Project’s Mission to End Suicide Among LGBTQ Youth’ (*The Business Wire*, 18 November 2019) <www.businesswire.com/news/home/20191118005423/en/PwC-Charitable-Foundation-Invests-6-Million-Trevor>.

¹⁴ Out and Equal, ‘Applying AI to LGBTQ Mental Wellness Support in the Workplace’ <<https://outandequal.org/wp-content/uploads/2018/10/Kimberley-Messer-Applying-AI-to-LGBTQ-mental-wellness-support-in-the-workplace.pdf>>.

¹⁵ KR Bisaso and others, ‘A Survey of Machine Learning Applications in HIV Clinical Research and Care’ (2017) 91 Computers in Biology and Medicine 366; JL Marcus and others, ‘Use of Electronic Health Record Data and Machine Learning to Identify Candidates for HIV Pre-Exposure Prophylaxis: A Modelling Study’ (2019) 10(6) *The Lancet HIV* 688.

¹⁶ JP Ridgway and others, ‘Machine Learning and Clinical Informatics for Improving HIV Care Continuum Outcomes’ (2021) 3 *Current HIV/AIDS Reports* 229.

¹⁷ See the chapter by Marília Papaléo Gagliardi in this volume. ‘Deepfake’ is a portmanteau of ‘deep learning’ and ‘fake’ and consists of synthetic media created by AI. After Chechen reports of

be deployed in the future more often to protect the safety and privacy of LGBTQ+ individuals, for instance, in depositions or court proceedings on LGBTQ+-related hate crime. Moreover, AI-powered technologies (such as certain social media) have facilitated the creation of digital (safe) spaces for the LGBTQ+ community to communicate, organise and express itself and to promote social wellness.¹⁸ In addition, ML in general can be used to mitigate censorship of LGBTQ+ issues or content; statistical data of deleted posts on LGBTQ+ content can be analysed to prevent future censorship and to hold LGBTQ+ rights abusers accountable through digital paper traces.¹⁹ An example of this was to be seen in research where chances of homophobia were predicted with an 89.4 per cent accuracy through ML and natural language analysis.²⁰ Deep learning algorithms have also been deployed to measure responses after the Supreme Court of India's decriminalisation of homosexuality.²¹ The biggest project in this field has been a partnership between GLAAD (Gay and Lesbian Alliance Against Defamation) and Google's parent company Alphabet and its subdivision Jigsaw. This partnership is focused on changing the way AI understands and interprets online LGBTQ+-related content. GLAAD has facilitated Google in training the AI that controls online algorithms, teaching it which phrases are offensive to the LGBTQ+ community and which are acceptable.²² Finally, a last example is the data collection by the city of San Francisco on the SOGI of LGBTQ+ San Franciscans through surveys on a range of demographics within the city government workforce.²³ The data is collected to analyse and track the progress made to ensure equity across many services such as healthcare, homeless, and housing services. There are also examples where big data can help promote LGBTQ+ rights.²⁴

the torture and detention of LGBTQ+ individuals, director David France embedded himself with Chechen LGBTQ+ activists to provide proof in his documentary *Welcome to Chechnya* (HBO, 2020) <welcometochechnya.com>. France used AI techniques to swap out faces of activists to protect their privacy, see KC Ifeanyi, 'How AI Came to Protect the LGBTQ Subjects in HBO's "Welcome to Chechnya"' (*Fast Company*, 30 June 2020) <[¹⁸ GLAAD, 'Social Media Safety Index' \(2021\) <\[https://glaad.org/sites/default/files/images/2021-05/GLAAD%20SOCIAL%20MEDIA%20SAFETY%20INDEX_0.pdf\]\(https://glaad.org/sites/default/files/images/2021-05/GLAAD%20SOCIAL%20MEDIA%20SAFETY%20INDEX_0.pdf\)>.](https://www.fastcompany.com/90522330/how-ai-came-to-protect-the-lgbtq-subjects-in-hbos>Welcome-to-chechnya>.</p>
</div>
<div data-bbox=)

¹⁹ A Bishop, 'No Access: LGBTIQ Website Censorship in Six Countries', Outright Action International (Report, 2021) <https://outrightinternational.org/sites/default/files/2022-09/NoAccess_abridged_1.pdf>.

²⁰ V Gomes Pereira, 'Using Supervised Machine Learning and Sentiment Analysis Techniques to Predict Homophobia in Portuguese Tweets' (PhD Dissertation, Fundação Getulio Vargas 2018).

²¹ A Khatua and others, 'Tweeting in Support of LGBT?: A Deep Learning Approach' (CoDS-COMAD 2019: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, January 2019) 342–45.

²² M Lamagna, 'How AI Could Make the Internet a Safer, More Civil Place' *New York Post* (14 March 2018) <<https://nypost.com/2018/03/14/how-ai-could-make-the-internet-a-more-civil-place/>>.

²³ San Francisco Administrative Code, Chapter 104: Collection of Sexual Orientation and Gender Identity Data.

²⁴ A Chowdry, 'How Big Data Can Further LGBTQIA+ Rights Beyond Pride Month' (*Quilt.AI*, 1 July 2021) <<https://quilt.ai/post/big-data-further-lgbtqi-rights-beyond-pride-month>>.

2.2 Negative Examples of AI Applications Relevant to the LGBTQ+ community

As with human rights in general, the use of AI is not entirely without potential harms or challenges for LGBTQ+-related rights. These can be sorted into roughly four categories which in practice often have overlap: censorship, privacy, health and safety, and equality and non-discrimination.

One of the main problematic implications of AI for the LGBTQ+ community is that it might lead to censorship. Censorship takes place intentionally when AI applications are programmed in such a way as to limit the freedom of expression and association of LGBTQ+ individuals and/or organisations through restrictions, censorship, or deletion of websites, apps, profiles, and so on. However, online content linked to marginalised groups (such as the LGBTQ+ community) tends to attract more negative feedback which, consequently, could also result in AI algorithms censoring content unintentionally. Minority groups such as the LGBTQ+ community are already relegated to the margins of many societies, thus not having access to AI technology or resources, in general, could lead to a further deterioration of their position. Furthermore, being forcefully ‘cut off’ is a direct violation of their fundamental rights such as the freedom of expression. Last year, Tencent’s WeChat (an instant messaging, social media, and mobile payment app commonly used in China), deactivated and deleted university-based LGBTQ+ groups and accounts in China.²⁵ Similar censoring has taken place in Russia under its anti-LGBTQ+ propaganda law,²⁶ but also in Indonesia, Malaysia, Iran, Saudi Arabia, and the United Arab Emirates (UAE).²⁷

Another category of AI use that needs extra caution and has multiple human rights implications (for privacy, but also for health and safety), is when sensitive SOGI data is collected and/or analysed. A prime example is a Stanford study that used deep neural networks to extract features from 35,326 facial images; these were then analysed by an AI algorithm developed with the aim to more accurately guess people’s sexual orientation.²⁸ With only a single facial image, the algorithm could correctly distinguish between gay and heterosexual men in 81 per cent of the cases; for women, this was 71 per cent. Humans were accurate only in 61 per

²⁵ N Gan, ‘WeChat Deletes Dozens of University LGBT Accounts in China’ (*CNN Business*, 9 July 2021) <<https://edition.cnn.com/2021/07/07/business/china-lgbt-wechat-censorship-intl-hnk/index.html>>. The students and associations indicated they provided safe spaces for LGBTQ+ individuals to communicate with each other and organise relevant events.

²⁶ N Wright (ed), *AI, China, Russia, and the Global Order: Technological, Political, Global, and Creative Perspectives* (Report of the United States Department of Defense 2018).

²⁷ Outright Action International (n 19).

²⁸ Y Wang and M Kosinski, ‘Deep Neural Networks are More Accurate than Humans at Detecting Sexual Orientation from Facial Images’ (2018) 114 *Journal of Personality and Social Psychology* 246–57. This study has later been criticised for its flawed data set, see A Burdick, ‘The AI “Gaydar” Study and the Real Dangers of Big Data’ *The New Yorker* (15 September 2017) <www.newyorker.com/news/daily-comment/the-ai-gaydar-study-and-the-real-dangers-of-big-data>.

cent of the cases for men and 54 per cent for women. The accuracy of the algorithm increased to 91 per cent and 83 per cent respectively when there were five facial images per person available, suggesting that the algorithm had ‘a better “gaydar” than humans.’²⁹ However, given that private actors and states are increasingly using AI applications to screen people’s intimate traits, the Stanford researchers emphasised the dangers of such developments for the privacy and safety of LGBTQ+ individuals. Another example where such technology was deployed on gender characteristics, is the ‘woman-only’ Giggle app. The Australian developed app has been taken to the Australian Human Rights Commission after using facial recognition software to prevent men who identify as women from joining.³⁰ Such AI applications and the data they collect can become dangerous in hands of individuals or companies with malicious intentions. They could, for instance, be used to reinforce ideologies based on eugenics.³¹ Considering there are seventy countries where homosexuality is still criminalised and eleven of these have the death penalty for it,³² AI techniques like these could be used by repressive regimes for heinous crimes against LGBTQ+ people (ie forced ‘outings’, conversion therapy, imprisonment, torture, other forms of violence, and even death).³³ There are already various examples of apps targeting the LGBTQ+ community specifically and offering conversion therapy services.³⁴

Finally, another category of concern for AI use is for equality and non-discrimination purposes. Minority groups as such are often already impacted or discriminated against more due to their intersectional qualities;³⁵ this can unintentionally be intensified in the application of AI.³⁶ AI can magnify existing inequalities because of certain biases that can unintentionally be programmed into

²⁹ S Levin, ‘LGBT Groups Denounce “Dangerous” AI that Uses Your Face to Guess Sexuality’ *The Guardian* (9 September 2017) <www.theguardian.com/world/2017/sep/08/ai-gay-gaydar-algorithm-facial-recognition-criticism-stanford>.

³⁰ L Bartlett, ‘Gender Wars: “Woman-Only” App in Trouble After Transgender Ban’ (*6PR*, 21 March 2022) <www.6pr.com.au/gender-wars-woman-only-app-in-trouble-after-transgender-ban>.

³¹ Tomasev and others (n 3).

³² Data retrieved from <www.zeroflagsproject.nl> and <<https://antigaylaws.org>>.

³³ See N Culzac, ‘Egypt’s Police Using Social Media and Apps like Grindr to Trap Gay People’ *Independent* (17 September 2014) <www.independent.co.uk/news/world/africa/egypt-s-police-using-social-media-and-apps-grindr-trap-gay-people-9738515.html>; T Fitzsimons, ‘Russian LGBTQ Activist is Killed After Being Listed on Gay-Hunting Website’ (*NBC News*, 23 July 2019) <www.nbcnews.com/feature/nbc-out/russian-lgbtq-activist-killed-after-being-listed-saw-inspired-site-n1032841>.

³⁴ R Ratcliffe, ‘Malaysian Government’s “Gay Conversion” App Pulled by Google Play’ *The Guardian* (17 March 2022) <www.theguardian.com/world/2022/mar/17/malaysian-governments-gay-conversion-app-pulled-by-google-play>; J Roettgers, ‘Under Pressure, Google Removes Gay Conversion Therapy App’ (*Reuters*, 30 March 2019) <www.reuters.com/article/variety-idUSL3N21G5E5>.

³⁵ Gaunt (n 11); European Commission DG for Justice and Consumers, *Report on Intersectional Discrimination in EU Gender Equality and Non-Discrimination Law* (EU, 2016).

³⁶ P Dhar, ‘A History of AI Bias: Achieving Algorithmic Justice for the LGBTQ Community’ (*Dell Technologies*, 30 June 2021) <www.delltechnologies.com/en-us/perspectives/history-ai-bias-lgbtq-community>; O Akselrod, ‘How Artificial Intelligence Can Deepen Racial and Economic Inequities’ (*ACLU*, 13 July 2021) <www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities>.

it and further perpetuate concerning ideas and beliefs on being *queer*.³⁷ While it is designed to fulfil human-defined goals, AI is likely to reflect the values and choices of the ones who build and use it.³⁸ Considering there are societal biases in the tech industry itself concerning the notions of sex, gender, and sexual orientation and a lack of LGBTQ+ representation,³⁹ some in academia, therefore, suggest that AI is set up to ‘always fail LGBTQ+ people’ as it is not developed with LGBTQ+ individuals in mind.⁴⁰ Simply put, AI programming, models, and techniques do not take the ‘queerness’ of people into account and the diversity that exists in the spectrum of gender, sexuality, and sexual orientation.⁴¹ This can have far-reaching consequences;⁴² the European Parliament (EP) has for instance stressed that many algorithmically driven identification technologies currently in use disproportionately misidentify and misclassify and therefore cause harm to LGBTQ+ people.⁴³

It is clear that both the negative and positive examples of AI applications reveal that the broad-scale state and commercial use of AI (like with any emerging technology) warrants the need for clear legal and ethical rules. These also need to be enforceable for accountability, responsibility, and liability reasons. The need for regulation thus grows with each day.

3 Regulatory Human Rights Framework and Protection from AI-Induced LGBTQ+ Discrimination

Data-driven technologies such as the internet and AI are not the subject of specific major international human rights law (IHRL) treaties; most of these instruments were adopted in the aftermath of World War II when such technologies

³⁷ See eg C Criado Perez, *Invisible Women. Exposing Data Bias in a World Designed for Men* (Vintage 2020); Tomasev and others (n 3) 3.

³⁸ Alan Turing Institute (n 1) 15.

³⁹ Dhar (n 36); J Manyika, J Silberg, and B Presten, ‘What Do We Do About the Biases in AI?’ (*Harvard Business Review*, 25 October 2019) <<https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai>>; N Sharkey, ‘The Impact of Gender and Race Bias in AI’ (*Humanitarian Law and Policy*, 28 August 2018) <<https://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/>>; A Kleinbort, ‘Gender and Artificial Intelligence: What Are the Current Issues and Challenges?’ (*Towards Data Science*, 8 February 2020) <<https://towardsdatascience.com/gender-and-artificial-intelligence-5fcff34589a>>; K Stathouopoulos and JC Mateos-Garcia, ‘Gender Diversity in AI Research’ (SSRN, 29 July 2019) <<https://ssrn.com/abstract=3428240>>.

⁴⁰ See the interview with anthropologist Mary L Gray in J Wareham, ‘Why Artificial Intelligence is Set Up to Fail LGBTQ People’ *Forbes* (21 March 2021) <www.forbes.com/sites/jamiewareham/2021/03/21/why-artificial-intelligence-will-always-fail-lgbtq-people/?sh=b3e657b301e7>.

⁴¹ D Leufer, ‘Computers Are Binary, People are Not: How AI Systems Undermine LGBTQ Identity’ (*Access Now*, 6 April 2021) <www.accessnow.org/how-ai-systems-undermine-lgbtq-identity/>.

⁴² M Carlisle, ‘A Queer Tax. New Lawsuit Alleges Aetna Discriminates Against LGBTQ People Seeking Fertility Treatment’ *Time* (14 September 2021) <<https://time.com/6097850/aetna-lgbtq-fertility-lawsuit/>>; *Goidel v Aetna Inc* Case No 1:21-cv-07619 Class Action Complaint (United States District Court Southern District Of New York, 2021).

⁴³ Point 9 of the European Parliament Resolution of 6 October 2021 on Artificial Intelligence in Criminal Law and its Use by the Police and Judicial Authorities in Criminal Matters (2020/2016(INI)).

did not exist or were not as developed as they are currently.⁴⁴ This does not entail that international law is not applicable. International human rights instruments⁴⁵ codify various types of fundamental rights which might be affected by the development, use, and application of AI; the scope or ambit of these instruments may be opened when AI touches upon provisions dealing with, *inter alia*, equality and non-discrimination, privacy, liberty and justice, civil and political rights and social and economic rights. Even if states have not ratified certain instruments and do not acknowledge the LGBTQ+ community, states are still not released from their obligations under international law. The fundamental rights contained in the aforementioned instruments are granted to all people, regardless of their SOGI, under customary international law which subsequently binds states as mandatory norms. States, therefore, are obligated to uphold fundamental rights and prevent violations; this includes the rights of LGBTQ+-oriented individuals or organisations. States also have the duty to protect against human rights abuses by third parties (including private actors) or where violations may be attributed to states.⁴⁶ Specific efforts on the European continent are being made by the Council of Europe (CoE)⁴⁷ and the European Union (EU)⁴⁸ to ensure that some regional regulatory framework on AI is in place.

⁴⁴ The same goes for LGBTQ+ rights; there are no international or regional treaties that specifically or exclusively address LGBTQ+ rights.

⁴⁵ Eg Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR). Note that the UDHR is not legally binding as an instrument and many of the rights reflected in it are enshrined in other legally binding instruments such as the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights. See International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR); International Covenant on Economic, Social and Cultural Rights 993 UNTS 3 (ICESCR).

⁴⁶ United Nations, 'Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework' (2011) 4 <https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf>.

⁴⁷ See eg the Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS 108). Furthermore, in April 2022, negotiations in the CoE started on an internationally binding legal instrument on AI, see <www.coe.int/en/web/artificial-intelligence/home>.

⁴⁸ See eg Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (General Data Protection Regulation/GDPR). Furthermore, the European Commission published an AI package in 2021, proposing new rules and actions for trustworthy AI which respects the EU fundamental rights and values. Part of the package was a proposal for an AI Act, see European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>. The EU has also adopted regulations in the form of a Digital Markets Act and a Digital Services Act which aim to create a safer digital space where the fundamental rights of users are protected and where a level playing field for businesses is established, see Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), OJ L 265, 12.10.2022, p. 1–66; and Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277, 27.10.2022, p. 1–102.

Private actors have the responsibility to adhere to the Guiding Principles on Business and Human Rights (UNGPs), developed by the United Nations (UN) to specifically address businesses and corporations.⁴⁹ SOGI are not mentioned therein, yet the Principles do highlight the responsibility of private actors to carry out human rights due diligence and to comply with human rights law (HRL).⁵⁰ Furthermore, states are encouraged to enact national laws to enable private actors to respect human rights, and are expected to enforce and periodically assess them.⁵¹ Still, the need for a legally binding instrument remains. A treaty on business and human rights has therefore been proposed, yet no consensus for adoption has been reached.⁵² Many private actors often already anticipate regulatory measures and adopt their own voluntary codes of conduct on the ethical development, use, and application of AI.⁵³

In the field of LGBTQ+ rights specifically, the Yogyakarta Principles addressing a broad range of international human rights standards and their application to SOGI issues, play an important role.⁵⁴ The Yogyakarta Principles are also not legally binding as such, yet do serve as an interpretive aid to other binding human rights instruments.

As the international or regional regulatory human rights framework hardly mentions or addresses SOGI-related rights that could specifically prevent AI applications from violating rights, a different approach is needed to protect from LGBTQ+ discrimination.⁵⁵ Section 5 of this chapter focuses on queer theory as a way to ensure that LGBTQ+ interests and values are considered from the development stages of AI on with the aim of this resulting in more (LGBTQ+) fairness in and by AI.

⁴⁹ Guiding Principles on Business and Human Rights (n 46).

⁵⁰ ibid 1.

⁵¹ ibid 4–6. Ironically, an unwanted result of the states' drive toward enacting legislation is excessive censorship and imposing data-collection requirements on the private sector, see Freedom House, 'Freedom on the Net 2021' (Report 2021) <https://freedomhouse.org/sites/default/files/2021-09/FOTN_2021_Complete_B booklet_09162021_FINAL_UPDATED.pdf>.

⁵² A Third Revised Draft of the Treaty has been published in August of 2021; full text available at <www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/WGTransCorp/Session6/LBI3 rdDRAFT.pdf>.

⁵³ Alan Turing Institute (n 1) 27.

⁵⁴ In 2017, additional principles expanding on the original document reflecting developments in IHRL and practice were published, The Yogyakarta Principles plus 10 (YP + 10). The new document also contains 111 'additional state obligations', related to areas such as torture, asylum, privacy, health, and the protection of human rights defenders. The full text of the Yogyakarta Principles plus 10 is available at <<https://yogyakartaprinciples.org>>.

⁵⁵ The Charter of Fundamental Rights of the European Union (the CFR) explicitly mentions sex and sexual orientation in its non-discrimination provision of Article 21, yet this provision is very generally worded. Furthermore, the CFR has only been binding since 2009, see European Union, Charter of Fundamental Rights of the European Union, 26 October 2012, 2012/C 326/02.

4 Queer Theory

So, how does *queer theory* come into play in this? Queer theory was initially introduced in the 1980s as a concept to critically reflect on the notions of gender, sexuality, sexual orientation, and the practices related to it. Inspired by the philosopher Michel Foucault who considered sexuality a cultural construct, queer theorists have developed the theory further from the 1990s onwards. Queer theory's primary concerns are to 'challenge binary approaches to sex and gender' and to 'offer critical reflection on the construction of sexual identities'.⁵⁶ To address these concerns, the technique of 'deconstruction' is employed. Deconstruction is a method to understand the relationship between text and its meaning.⁵⁷ It focuses on the notion that language is based on a system full of hidden antinomies, hierarchy, and oppositions. Language is therefore considered a system in which words exist because of being in opposition to others and in which some words govern over others. Deconstruction in the framework of queer theory entails an analysis of the traditional views on gender, sexuality, and sexual orientation that determine how society looks at these notions (ie through the dominant binary discourse on the dichotomies of male/female, masculine/feminine, heteronormative/homonormative, cisgender/transgender, and so on), followed by a temporary reversal of the hierarchy and the active offering of different or multiple views on how these notions can be interpreted.⁵⁸ Applying deconstruction to AI and *queerifying* it would be looking at how these notions are programmed in or interpreted by AI, reversing the hierarchy and offering *queer* options; this will be demonstrated in section 5 with three case studies. The benefits of a *queerification* of AI are that it may help to expose any underpinnings, assumptions, beliefs, and/or biases that underlie the interpretation of the aforementioned notions in AI programming and applications; a *queerification* of AI could lead to it becoming more diverse, inclusive and safe for the LGBTQ+ community.

5 A Queerification of AI

Let us consider three case studies in which AI is *queerified*. Take the example of gender for instance: most individuals regularly have to fill out online applications and forms in which they are asked to indicate their gender. Oftentimes it is simply for data collection or statistics without any specific (legal) reasons or needs why

⁵⁶ D Gonzalez-Salzberg and L Hodson (eds), *Research Methods for International Human Rights Law: Beyond the Traditional Paradigm* (Routledge 2020).

⁵⁷ Deconstruction is derived mainly from the works of linguist Ferdinand De Saussure and philosopher Jacques Derrida, see J Balkin, 'Deconstructive Practice and Legal Theory' (1987) 96(4) Yale Law Journal 743.

⁵⁸ *ibid* 746.

one is asked about their gender. The choices provided to pick a gender are usually limited to ‘male’ and ‘female’.⁵⁹ Research conducted with data spanning five decades and from seventeen countries, estimates that up to 2 per cent of the general world population identifies as transgender or gender nonconforming (think of non-binary, genderqueer, and so on) depending on the inclusion criteria and geographic location.⁶⁰ Based on the current growth of the world population, this would mean that up to almost 160 million people worldwide do not necessarily identify their gender as male or female; a number as large as roughly half of the population of the United States (US).⁶¹ There are different options for making the AI behind applications and forms more inclusive for these individuals. One option would be to not use the dominant binary model of ‘male’ and ‘female’ and swap it for a binary ‘plus’ model including a third gender option of ‘other’ or ‘prefer not to answer’. This model has however also been critiqued for that it may alienate people with the option of ‘other’. The best option would be to use the ‘user input model’⁶² which has certain fixed gender options and an additional empty user input box with the option for individuals for ‘auto-identification’ of their gender as they see it.⁶³ ‘Auto-identification’ has been acknowledged by the Inter-American Court of Human Rights in its Advisory Opinion on ‘gender identity, and equality and non-discrimination of same-sex couples’ as a right.⁶⁴ Other courts have established that ‘misgendering’ individuals constitutes discrimination.⁶⁵

A second case study on *queerifying* AI is on sexual orientation. Recent American research revealed that one in six American adults in ‘Generation Z’ consider themselves to be part of the LGBTQ+ community as regards sexual orientation.⁶⁶ Yet, the notion of sexual orientation is programmed into many AI forms, programs, and software without taking the full spectrum of the diversity of the sexual orientations of people into consideration. This was recently highlighted in the case of insurance company Aetna that had denied a lesbian couple intrauterine insemination (IUI)

⁵⁹ O Keyes, ‘The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition’ (2018) 2 Proceedings of the ACM on Human-Computer Interaction 88.

⁶⁰ M Goodman and others, ‘Size and Distribution of Transgender and Gender Nonconforming Populations: A Narrative Review’ (2019) 48(2) Endocrinology and Metabolism Clinics of North America 303.

⁶¹ In the US alone, there are around 1.2 million individuals that identify as non-binary, see <<https://williamsinstitute.law.ucla.edu/publications/nonbinary-lgbtq-adults-us/>>.

⁶² E Shachar-Hill, ‘A Guide to Creating LGBTQ-Inclusive Forms’ (*Keshet*, 22 April 2020) <www.keshetonline.org/resources/a-guide-to-creating-lgbtq-inclusive-forms/>.

⁶³ A Krzyminski, ‘The Complex UX Design Behind Gender Selection Forms’ (*UX Collective*, 4 May 2020) <<https://uxdesign.cc/the-complex-ux-design-behind-gender-selection-forms-ddfa298e2bb>>. To prevent people filling out unrelated terms, some entry verification conditions are obviously needed to limit the potential responses.

⁶⁴ *State Obligations Concerning the Change of Name, Gender Identity, and Rights Derived from a Relationship Between Same-Sex Couples*, Advisory Opinion OC-24/17 (Inter-American Court of Human Rights, 24 November 2017), paras 115–16.

⁶⁵ *Nelson v Goodberry Restaurant Group Ltd dba Buono Osteria* [2021] BCHRT 137.

⁶⁶ S Schmidt, ‘1 in 6 Gen Z Adults Are LGBT. And this Number Could Continue to Grow’ *The Washington Post* (24 February 2021) <www.washingtonpost.com/dc-md-va/2021/02/24/gen-z-lgbt/>.

treatment for not meeting its heteronormative definition of infertility.⁶⁷ Aetna's policy stated that a person is considered infertile if he or she is unable to conceive or produce conception after one year of frequent, unprotected *heterosexual* sexual intercourse (or six months in case the female partner is thirty-five years or older). Outside of these options, persons were supposed to pay for twelve or six months of IUI, depending on their age. Considering many insurance companies and other organisations work with AI to pre-screen submitted claims, cases such as *Aetna* demonstrate how unintentional heteronormative programming impacts the LGBTQ+ community significantly. *Queerifying* AI in this case would entail programming sexual orientation in more inclusive terms by using LGBTQ+ semantics instead of the dominant heteronormative options.

The third case study relates to the fairness of AI for the LGBTQ+ community. Fairness can be achieved in different ways. One way is through its development and programming. As established in section 2.2 of this chapter, AI reflects the values of those developing, programming, and applying it. *Queerifying* AI and making it free of LGBTQ+ biases requires having diversity within the people and organisations behind the AI and including the LGBTQ+ community from the development stages on. This way, algorithms and ML models are programmed to be more objective early on. Another way of achieving fairness is through having a diverse and inclusive data set. There is a lack of data representing the LGBTQ+ community and of marginalised groups in general. LGBTQ+ data collection is not without risks, but when these are adequately calculated, considered, and addressed (and human rights violations are prevented), the benefits are tremendous.⁶⁸ *Queerifying* AI entails actively and regularly identifying where there are gaps in data and amplifying or boosting specific LGBTQ+ data inputs with preferably as intersectional data as possible;⁶⁹ partnering with organisations that have, for decades, been collecting LGBTQ+ data while taking fundamental rights of such individuals into account, such as the Trevor Project, the Williams Institute, ILGA and OutRight Action International, would facilitate this process.⁷⁰ Finally, a last way of achieving more LGBTQ+ fairness in AI is by measuring it. Many AI applications have a final 'filter' checking racial or gender bias. *Queerifying* AI would entail adding a 'LGBTQ+'

⁶⁷ *Goidel* (n 42).

⁶⁸ San Francisco Administrative Code (n 23)—the San Francisco model is a good example.

⁶⁹ Gaunt (n 11).

⁷⁰ The Trevor Project monitors, analyses, and evaluates data collected from LGBTQ+ young people to prevent LGBTQ+ youth suicide, see <thetrevorproject.org>; the Williams Institute is a think tank affiliated with the UCLA School of Law that produces multidisciplinary research to provide a foundation for laws and policies shaping the lives of LGBTQ+ people, see <williamsinstitute.law.ucla.edu>; ILGA is a worldwide federation of more than 1,700 organisations from over 160 countries and territories campaigning for LGBTQ+ human rights, see <ilga.org>; OutRight Action International ensures human rights for LGBTIQ people globally through strategic programme areas, see <outrightinternational.org>.

filter in many AI applications to check whether this specific group is being underserved or negatively impacted in any other way.⁷¹

6 Conclusion

This chapter has discussed examples of AI applications that are helping the LGBTQ+ community overcome many of the problems it faces in daily life. However, the negative examples of AI use reveal that the impact on human rights, especially in terms of privacy, discrimination, freedom of expression, among other areas, are enormous and underserve the LGBTQ+ community significantly, making it currently more of a foe than a friend. It is safe to say that the notions of gender, sexuality, and sexual orientation and the diversity therein currently cannot be grasped correctly by AI that is programmed to ‘think’ in binaries such as 0s and 1s, but also still in terms of male/female, masculine/feminine, heteronormative/homonormative, cisgender/transgender, and so on. This chapter has advocated the use of queer theory and deconstruction with a view towards helping expose any biases and assumptions that underlie AI programming that are detrimental for the LGBTQ+ community and has provided multiple examples of how a different *queer* approach would be fairer. Indeed, *queerifying* AI could lead it to being more inclusive and diverse for the LGBTQ+ community with the use of more relevant data sets, LGBTQ+ semantics, and LGBTQ+ ‘filters’ to tackle LGBTQ+ biases and an underserving of the community. A *queerification* of AI is not only needed, but also long overdue.

⁷¹ D Pope, ‘Five Ways to Bring a UX Lens to Your AI Project’ (*TechCrunch*, 21 July 2020) <techcrunch.com/2020/07/21/five-ways-to-bring-a-ux-lens-to-your-ai-project/>.

16

Artificial Intelligence and Women's Rights

Deepfake Technology

Marília Papaléo Gagliardi

1 Introduction

Technology is inextricably linked to the entertainment industry, which uses various innovations and digital techniques to improve its products, including computer-generated images.¹ Aiming to create better and more developed edits, synchronising the actors' speech and voice acting², as well as increasing the realism of the scene, tools that use deepfake AI have also been employed in the film industry³.

The use of deepfake for altering images is not just restricted to movies. Algorithms for creating and editing videos have also been commercialised in apps that allow the alteration of an original video⁴ by replacing someone's face,⁵ adding the face of another person,⁶ or transferring users' faces to movie characters.⁷ This form of entertainment presents an inherent risk: the possible creation of vexatious, shameful, or vulnerable contexts for those whose images were manipulated.

¹ *The Irishman* (Netflix) and *Rogue One: A Star Wars Story* used computer-generated imagery (CGI), as declared by the creators. See Dave Itzkoff, 'How "Rogue One" Brought Back Familiar Faces' *The New York Times* (27 December 2016) <www.nytimes.com/2016/12/27/movies/how-rogue-one-brought-back-grand-moff-tarkin.html>.

² The Flawless company created the system called TrueSync: '[A] software that uses deepfakes to synchronize the actors' mouths with the dubbing sound. The world's first system that uses Artificial Intelligence to create perfectly lip-synced visualisations in multiple languages.' See Flawless, 'Neural Network Enabled Filmmaking: Our Product' (*Flawless*, 2021) <www.flawlessai.com/product>.

³ In June 2020, Disney Research Studios presented an algorithm for fully automatic neural face-swapping in images and videos, which is nothing more than a deepfake technique. See <<<REFO:JART>>> Jacek Naruniec et al, 'High-Resolution Neural Face Swapping for Visual Effects' (2020) 39(4) Computer Graphics Forum 173.

⁴ Eg TikTok is a platform that allows for a series of changes in the backgrounds, sounds, and images as provided by the platform. See 'Termos de Serviço' (*TikTok*, July 2020) <www.tiktok.com/legal/terms-of-service?lang=pt_BR>.

⁵ Such as DeepFaceLab as provided by iperov. See 'DeepFaceLab: the leading software for creating deepfakes' (*GitHub*, 2021) <<https://github.com/iperov/DeepFaceLab>>.

⁶ Such as Impressions as provided by Helito Beggiora, 'Impressions App: Como Usar o Aplicativo de Deepfake' (*TechTudo*, 2 December 2020) <www.techtudo.com.br/dicas-e-tutoriais/2020/12/impressions-app-como-usar-o-aplicativo-de-deepfake.ghtml>.

⁷ Such as ZAO Deepfake as provided by Ana Letícia Loubak, 'Aplicativo Zao usa deepfake para criar vídeos e viraliza na China' (*TechTudo*, 3 September 2019) <www.techtudo.com.br/noticias/2019/09/aplicativo-zao-usa-deepfake-para-criar-videos-e-viraliza-na-china.ghtml>.

This situation can affect anyone, especially women, which can be inferred by a survey conducted by Sensity AI (previously Deeptrace Labs) investigating the willingness for editing technology to 'undress' a person's images. Most survey participants reported desiring to use deepfake techniques on women they know in real life (63 per cent). Only 6 per cent of respondents reported not being willing to use deepfake for this purpose.⁸

The motivations behind this desire to create deepfake content vary significantly. In the case of celebrity deepfakes, the creation can occur to both sexually objectify and construct a parasocial intimacy with these personalities.⁹ This would imply that users transgress the limits of respect for private life, violating the victim's right to privacy and intimacy.¹⁰ Indeed, some famous actresses have been victims of deepfake alterations that create fake pornography.¹¹

However, this issue can affect anyone, including anonymous women.¹² Thus, it is important to analyse how this technology has been used in the context of violence against women and what the human rights ramifications for the state and/or developers of these apps are. To do so, this chapter examines how the use of deepfake can intensify forms of gender discrimination, by studying cases and examining what potentially motivates the violence (section 2). It also analyses how this technology operates and why it can be harmful to women (section 3). Finally, it considers if there is any legal framework that encompasses the use of this technology, and potential mitigation strategies to remedy any misuse (section 4). The conclusion (section 5) relies on the fact that some measures must be taken so deepfake theology does not lead to any harm.

⁸ Eric Kocsis, 'Deepfakes, Shallowfakes, and the Need for a Private Right of Action' (2022) 126(2) Dickinson Law Review 621; Karen Hao, 'A Deepfake Bot Is Being Used to "Undress" Underage Girls' (*MIT Technology Review*, 20 October 2020) <www.technologyreview.com/2020/10/20/1010789/ai-deepfake-bot-undresses-women-and-underage-girls/>.

⁹ Milena Popova, 'Reading Out of Context: Pornographic Deepfakes, Celebrity and Intimacy' (2019) 7 *Porn Studies* 367.

¹⁰ Sophie Maddocks, '"A Deepfake Porn Plot Intended to Silence Me": Exploring Continuities Between Pornographic and "Political" Deep Fakes' (2020) 7 *Porn Studies* 415.

¹¹ For instance, this happened to Gal Gadot, Masiel Williams, and Daisy Ridley as mentioned in Russell Spivak, '"Deepfakes": The Newest Way To Commit One of the Oldest Crimes' (2019) 3(2) *Georgetown Law Technology Review* 339. It also happened to Taylor Swift, as exposed in Víctor Cerdán Martínez and Graciela Padilla Castillo, 'Historia del Fake Audiovisual: Deepfake y la Mujer en un Imaginario Falsificado y Perverso' (2019) 24(2) *Historia y Comunicación Social* 505; and to Scarlett Johansson as mentioned in Luciano Floridi, 'Artificial Intelligence, Deepfakes and a Future of Ectypes' (2018) 31 *Philosophy & Technology* 317.

¹² Chandell E Gosse and Jacquelyn Burkell, 'Politics and Porn: How News Media Characterizes Problems Presented by Deepfakes' (2020) 37(5) *Critical Studies in Media Communication* 497.

2 How the Use of Deepfake Can Intensify Forms of Gender Discrimination

Deepfake can use anyone's face and any video to perform alterations and montages, regardless of consent. As a result, the individual cannot determine 'what information about himself or herself should be known to others,'¹³ thus rendering impossible any control over personal information.¹⁴ However, the discussion about deepfake represents not only a privacy issue but also a gender issue,¹⁵ considering the sheer fact that 96 per cent of deepfake videos are pornographic and target women.¹⁶

Indeed, the 'undressing' of people online, without their knowledge, 'fetishizes the non-consent'.¹⁷ In that sense, under the lens of the cultural pattern of objectification of women's bodies,¹⁸ the use of deepfake represents a power fantasy stimulus for image editors. It is noteworthy that when considering the non-consensual dissemination of intimate images, this objectification of the female body can also imply a means by which hegemonic masculinity is portrayed and heterosexuality affirmed.¹⁹ This type of masculinity creates a hierarchy, legitimising a dominant position for men in society, while subordinating women,²⁰ reinforcing a pattern of gender oppression.

This pattern causes harm. Indeed, deepfake usage for porn can be considered a form of 'technology-facilitated sexual violence', which reinforces and maintains gender inequalities through harmful sexually aggressive and harassing behaviour perpetrated with technologies.²¹ It is specifically considered by legal scholars as an invasion of sexual privacy,²² and non-consensual pornography.²³

The non-consensual creation of private images can develop in different ways, which are constantly changing, due to technological advances. Since existing laws

¹³ Alan Westin, 'Social and Political Dimensions of Privacy' (2003) 59 *Journal of Social Issues* 431.

¹⁴ Daniel Solove, *Nothing to Hide: The False Tradeoff between Privacy and Security* (1st edn, Yale UP 2011).

¹⁵ Emily van der Nagel, 'Verifying Images: Deepfakes, Control, and Consent' (2020) 7(4) *Porn Studies* 424; Nicola Henry and Alice Witt, 'Governing Image-Based Sexual Abuse: Digital Platform Policies, Tools, and Practices' in Jane Bailey and others (eds), *The Emerald International Handbook of Technology-Facilitated Violence and Abuse* (Emerald 2021) 179.

¹⁶ Henry Ajder and others, 'The State of Deepfakes Landscape, Threats, And Impact' (*Deeptrace*, 2019) <https://regmedia.co.uk/2019/10/08/deepfake_report.pdf>.

¹⁷ Maddocks (n 10) 423.

¹⁸ Travis L Wagner and Ashley Blewer, '"The Word Real Is No Longer Real": Deepfakes, Gender, and the Challenges of AI-Altered Video' (2019) 3 *Open Information Science* 36.

¹⁹ Nathian Shae Rodriguez and Terri Hernandez, 'Dibs on that Sexy Piece of Ass: Hegemonic Masculinity on TFM Girls Instagram Account' (2018) 4(1) *Social Media and Society* 1.

²⁰ Raewyn Connell and James W Messerschmidt, 'Hegemonic Masculinity: Rethinking the Concept' (2005) 19(6) *Gender and Society* 829.

²¹ Henry Nicola and Anastasia Powell, 'Technology-Facilitated Sexual Violence: A Literature Review of Empirical Research' (2018) 19(2) *Trauma, Violence and Abuse* 195.

²² Danielle Citron, 'Sexual Privacy' (2019) 128 *Yale Law Journal* 1870.

²³ Mary Anne Franks, '"Revenge Porn" Reform: A View from the Front Lines' (2017) 69 *Florida Law Review* 1251.

do not always contemplate all forms of this violation,²⁴ the term ‘image-based sexual abuse’—which encompasses all these aggressions generated using sexual images²⁵—can also be applied.

In addition to the violation of women’s intimacy, this also evokes a non-consensual form of sexual exploitation, leading to a continued mediation and objectification of feminine and femme bodies.²⁶ This practice can cause mental and physical pain to victims. It can also be conceptualised as being in a ‘continuum of sexual violence’,²⁷ along with other forms of disrespect for consent, such as catcalling or rape, with the characteristic presence of misogyny despite affecting all genders. This kind of violence is synonymous with the ‘logic of rape’, a rationalisation and practice of abusive attitudes within the context of rape culture,²⁸ which can also be manifest through the use of online media.²⁹

It is also possible to acknowledge the discriminatory use of deepfake in the political sphere.³⁰ This happens when deepfake is used to stigmatise, shame, and expose women in positions of power in an attempt to silence their political positions,³¹ as a form of ‘retaliation’.³² Investigative journalist Rana Ayyub was a victim of pornographic deepfake after taking a stand against a context of sexual violence.³³ Hillary Clinton had fake sex videos of her leaked and promoted by an agency linked to Russia during the election period of 2016.³⁴ According to Maddocks, although political deepfakes are perceived as those capable of impacting democracy, pornographic deepfakes do also threaten democratic principles.³⁵ The author further argues that fake pornography seems to be applied specifically as a weapon against women, intending to impose upon them a position of disrepute.

²⁴ Liz Kelly, *Surviving Sexual Violence* (1st edn, University of Minnesota Press 1988).

²⁵ Clare McGlynn and Erika Rackley, ‘Image-Based Sexual Abuse’ (2017) 37(3) Oxford Journal of Legal Studies 534.

²⁶ Wagner and Blewer (n 18) 36; and Nicola Henry and Anastasia Powell, ‘Embodied Harms: Gender, Shame, and Technology-Facilitated Sexual Violence’ (2015) 21(6) Violence Against Women 758.

²⁷ Kelly (n 24) 76.

²⁸ Mary Anne Franks, ‘An-Aesthetic Theory: Adorno, Sexuality, and Memory’ in Renée J Heberle (ed), *Feminist Interpretations of Theodor Adorno* (Pennsylvania State UP 2006) 193.

²⁹ Despoina Mantzari, ‘Sadistic Scopophilia in Contemporary Rape Culture: I Spit On Your Grave and the Practice of “Media Rape”’ (2018) 18(3) Feminist Media Studies 397.

³⁰ Chenxi Wang, ‘Deepfakes, Revenge Porn, and the Impact on Women’ (*Forbes*, 1 November 2019) <www.forbes.com/sites/chenxiwang/2019/11/01/deepfakes-revenge-porn-and-the-impact-on-women>.

³¹ Maddocks (n 10) 418; Miosotis Soto Santana, ‘Justice for Women: Deep Fakes and Revenge Porn’ (3rd Global Conference on Women’s Studies, Rotterdam, 25–27 February 2022) <www.dpublication.com/wp-content/uploads/2022/02/27-10177.pdf>.

³² Bella Thorne, for example, had her images transposed to an explicit video created after fighting for causes against sexual violence, as mentioned in Maddocks (n 10) 416–17.

³³ Shannon Reid, ‘The Deepfake Dilemma: Reconciling Privacy and First Amendment Protections’ (2021) 23(1) University of Pennsylvania Journal of Constitutional Law 209, 210.

³⁴ Ben Collins, ‘Russia-linked Account Pushed Fake Hillary Clinton Sex Video’ (*NBC News*, 10 April 2018) <www.nbcnews.com/tech/security/russia-linked-account-pushed-fake-hillary-clinton-sex-video-n864871>.

³⁵ Maddocks (n 10) 422.

Despite the different forms of classifications and definitions applicable to the usage of deepfake for porn, and despite the various possible motivations of the editors that create such content, it is possible to say that this violence is gendered since it is used to control (predominantly) women.³⁶ Thus, the use of deepfake for this aim does not only imply a violation of privacy but also a form of the perpetuation of gender violence. The question that emerges is how this technology—the creation, production, and distribution of these videos and photos—operates and, therefore, allows the perpetration of this violence.

3 How Deepfake's Technology Operates and Why It Can Be Harmful

The harm caused by the use of deepfake AI relates, in part, to the images' realism, which derives from the deep learning (DL) tools used in the creation of deepfake. These tools allow computers to identify patterns and reproduce them quickly, generating new meanings from those patterns. The technological process involved in the creation and development of this technology allows deepfake to be identified as 'a prototype of Artificial Intelligence'.³⁷

It is important to highlight that the origin of the term 'deepfake' is connected to pornographic videos on the social network Reddit, published by a user with the nickname 'deepfakes' in 2017. This user created fake videos and placed the faces of Hollywood actresses on the bodies of the porn stars that were originally in the videos,³⁸ with very convincing results.

The similarity generated, in turn, stems from an algorithmic combination arising from the generative adversarial networks technology. This technology consists of two different algorithms: one that generates artificial images based on a database containing real images of the targeted model (generator), and the other that detects false images (discriminator).³⁹ Once this combination undergoes continuous training, the computer starts to create realistic images and videos, since each of the algorithms improves the operation of the other.⁴⁰ Eventually, the algorithm will generate a very convincing video.⁴¹

³⁶ Kelly (n 24) 76; and Nicola Henry, Asher Flynn, and Anastasia Powell, 'Policing Image-based Sexual Abuse: Stakeholder Perspectives' (2018) 19(6) Police Practice and Research 565.

³⁷ Wagner and Blewer (n 18) 36.

³⁸ *ibid* 32.

³⁹ Miha Šepc M and Melanija Lango, 'Virtual Revenge Pornography As A New Online Threat To Sexual Integrity' (2020) 15 Balkan Social Science Review 117.

⁴⁰ Mika Westerlund, 'The Emergence of Deepfake Technology: A Review' (2019) 9 Technology Innovation Management Review 40.

⁴¹ Spivak (n 11).

In this context, deepfake technology has improved over time, to the point that the entire action of altering a person's image can derive from a single image.⁴² The danger of violation of privacy or non-discrimination, therefore, becomes even higher, since the risk can arise from a single photo. This medium, it is important to reiterate, can be available on the internet regardless of a victim's will.

Deepfake is used in different forms. In June 2019, the existence of an app called DeepNude was reported, which used AI (deepfake) to 'undress' women.⁴³ The app allowed for the editing of a photo (even with a low-quality image) of a fully clothed woman, into an image showing the same woman naked. Interestingly, this app did not work on men. After complaints, the app's creators quickly removed it from circulation⁴⁴ and announced that they would not commercialise the technology.

Very similar technology was also being used by a publicly available bot in the Telegram messaging app. It allowed users to upload a photo through the app (mobile or web-based) and to receive a nude picture back within minutes. The reach of this bot was gigantic, and, by July 2020 at least, 100,000 women were already 'undressed'. The victims' ages varied widely and many underage girls were victims of deepfake photos, too.⁴⁵ After the publication of the research reporting these practices, the bot was blocked on iOS for violating App Store guidelines.⁴⁶ The main Telegram channel that hosted the bot, as well as an affiliate channel to share its creations, have now been removed.

Accordingly, these technological advancements may cause serious harm on victims. Even when it is possible to retroactively identify that a picture being circulated is an altered image, the reputational damage has already been done. The reality is that it does not matter if the images were 'originally' sexual or actually constructed by an app. The Australian Women Against Violence Alliance states that 'the fact that an image has been altered, or even composed of images taken of different women, does not lessen the potential harm resulting from its dissemination'.⁴⁷ And, even without high accuracy, the creation of harmful video achieves its purpose.

⁴² Joan Solsman, 'Samsung Deepfake AI Could Fabricate a Video of You From a Single Profile Pic' (CNET, 24 May 2019) <www.cnet.com/tech/computing/samsung-ai-deepfake-can-fabricate-a-video-of-you-from-a-single-photo-mona-lisa-cheapfake-dumbfake/>.

⁴³ Anne Pechenik Gieseke, "'The New Weapon of Choice': Law's Current Inability to Properly Address Deepfake Pornography" (2020) 73(5) Vanderbilt Law Review 1479, 1487.

⁴⁴ Samantha Cole, 'Creator of DeepNude, App That Undresses Photos of Women, Takes It Offline' (*Vice*, 27 June 2019) <www.vice.com/en/article/qv7agw/deepnude-app-that-undresses-photos-of-women-takes-it-offline>.

⁴⁵ Hao (n 8).

⁴⁶ According to the researchers, on their Twitter page. See Henry Ajder, 'Breaking: It Appears the Telegram Deepfake Bot Has Been Blocked on iOS for Violating App Store Guidelines' (*Twitter*, 28 October 2020) <<https://mobile.twitter.com/HenryAjder/status/1321529152009359362>>.

⁴⁷ Australian Women Against Violence Alliance (AWAVA), 'Submission to the Senate Inquiry into the Phenomenon Colloquially Referred to as "Revenge Porn"' (AWAVA, 14 January 2016) <<https://awava.org.au/wp-content/uploads/2016/04/AWAVA-submission-to-Senate-inquiry-on-phenomenon-colloquially-referred-to-as-revenge-porn-Jan-2016.pdf>>.

Violations arise, therefore, from the use of a technology that tends to become more and more specialised. For this reason, it is important to regulate the use of deepfake, aiming to mitigate the damage generated, and to hold the perpetrators of such violence accountable. This, however, can become a tall order given the lack of applicable provisions in the current regulatory framework. Section 4 will discuss this aspect.

4 Current Regulatory Framework and Recommendations

From an international law perspective, once states have signed and ratified international human rights treaties, they have a duty to respect, protect and fulfil these rights.⁴⁸ This means that it is not only up to states to refrain from violating human rights within their territory but to act proactively to protect them and ensure that they are being preserved and achieved.

Although the international treaties do not contain, in their articles, a specific reference to violations generated using deepfake, the right to privacy⁴⁹ is internationally recognised and so is the protection against gender-based violence.⁵⁰ Those human rights standards apply to the violations perpetrated with the use of technology since rights that people have offline must also be protected online.⁵¹ In other words, the current international law is applicable to situations of the use of pornographic deepfake without the consent of the victims.

That said, even though many states have committed themselves internationally and domestically to the right of privacy and to the duty to prevent gender-based discrimination, very little has been done to protect, promote, and fulfil them when it comes to deepfake abuses.

Notably, regarding the legislative sphere, laws at the national level that recognise and condemn gendered violence created with the use of deepfake are very scarce indeed.⁵² The applications of the existing law, in turn, do not properly address the

⁴⁸ David Jason Karp, ‘What is the Responsibility to Respect Human Rights? Reconsidering the “Respect, Protect, and Fulfill” Framework’ (2020) 12(1) *International Theory* 83.

⁴⁹ The right to privacy is protected under article 12 of the Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR). The right to private life is also protected in article 17 of the International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR).

⁵⁰ Provisions regarding the prohibition of discrimination against women are present in art 7 of the UDHR and the entire Convention on the Elimination of All Forms of Discrimination against Women (adopted 18 December 1979, entered into force 3 September 1981) 1249 UNTS 13 (CEDAW).

⁵¹ Eg Human Rights Council, ‘The Right to Privacy in the Digital Age’ (7 October 2019) UN Doc A/HRC/RES/42/15.

⁵² Eg in the United States (US), the only states that explicitly prohibit deepfake media are Virginia and California. Even in places that regulate revenge porn, like the United Kingdom (UK), the law does not encompass fake or edited images, as observed by Karolina Mania, ‘The Legal Implications and Remedies Concerning Revenge Porn and Fake Porn: A Common Law Perspective’ (2020) 24 *Sexuality & Culture* 2079; see also Karen Hao, ‘Deepfake Porn is Ruining Women’s Lives. Now the Law May

damage caused. For example, criminal 'defamation' offences may cover the use of images to offend another person. However, such an offence typically relies on the intent of the creator of the videos to cause harm,⁵³ which is not always the case. Although sometimes there is a motivation to generate embarrassment or a form of retaliation to harm women, defamation and other laws that require 'specific intent' fail to address cases where the primary use of this technology is financial gain, sexual gratification, or entertainment.⁵⁴ As a consequence, the application of the existing laws may be blocked or problematic in many cases.

In addition, the use of other legal standards that also rely on intent could also generate uncertainty about enforcement since they may depend on causing 'serious harm'.⁵⁵ The victim would have to prove the extent and the consequences of the violation, thus being exposed to revictimisation.⁵⁶

Domestic courts do not present a proper avenue to these human rights violations in cases of deepfake either. The primary reason being that, to present a case to a court, there must be someone to be accused of disrespecting a right. Deepfake content, however, might originate from anonymous sources. As regards many internet and technology-based crimes, the platforms that reproduce the content are clearly identifiable, whereas the creators/perpetrators of the actual crimes are not.⁵⁷ This means that, in addition to being impossible to hold individuals responsible for their actions and punish them for a criminal offence, it is also difficult for the victims to secure any financial reparation. Therefore, neither the victims nor the actresses of the original films which hold copyright may be able to claim financial reparation for a specific violation.

Even when the deepfake does not originate from an anonymous source, or when investigations would identify who is responsible, other problems might arise since

'Finally Ban It' (*MIT Technology Review*, 12 February 2021) <www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/>.

⁵³ Van Vechten Veeder, 'The History and Theory of the Law of Defamation: II' (1904) 4(1) *Columbia Law Review* 33.

⁵⁴ Henry, Flynn, and Powell (n 36) 568; Philip Hayward and Alison Rahn, 'Opening Pandora's Box: Pleasure, Consent and Consequence in the Production and Circulation of Celebrity Sex Videos' (2015) 2(1) *Porn Studies* 1; Katherine Gabriel, 'Feminist Revenge: Seeking Justice for Victims of Nonconsensual Pornography through "Revenge Porn" Reform' (2019) 44 *Vermont Law Review* 849.

⁵⁵ Tom Rudkin, 'Things Get Serious: Defining Defamation' (2014) 25(6) *Entertainment Law Review* 201.

⁵⁶ Dale Spencer and others, '"I Think It's Re-Victimizing Victims Almost Every Time": Police Perceptions of Criminal Justice Responses to Sexual Violence' (2018) 26(2) *Critical Criminology* 189; Cheryl Regehr and Ramona Alaggia, 'Perspectives of Justice for Victims of Sexual Violence' (2006) 1(1) *Victims and Offenders* 33; Kaitlyn Regehr, Arijah Birze, and Cheryl Regehr, 'Technology Facilitated Re-Victimization: How Video Evidence of Sexual Violence Contributes to Mediated Cycles of Abuse' (2021) 18(4) *Crime Media Culture* 1.

⁵⁷ Wagner and Blewer (n 18) 38; Henry, Flynn, and Powell (n 36) 571; Roderic Broadhurst, 'Developments in the Global Law Enforcement of Cyber-Crime' (2006) 29(3) *Policing: An International Journal of Police Strategies & Management* 408; Cameron SD Brown, 'Investigating and Prosecuting Cybercrime: Forensic Dependencies and Barriers to Justice' (2015) 9(1) *International Journal of Cyber Criminology* 55.

legal prosecution demands significant time and financial resources, while there also remains the possibility that the perpetrator is not able to pay for the damage caused.⁵⁸

Even in a scenario in which the perpetrator is brought to court and is appropriately held accountable, there could also be still further problems regarding the amount of compensation. Although there are convictions for moral damages and psychological damages, those cases have historically assigned lower values than in the context of readily quantifiable economic damage, unveiling further limitations to the enforcement of the right to privacy.⁵⁹

Another limitation regarding court proceedings might be the possible international component of deepfake-based privacy and gender-violence crimes. Legal remedies are not always applicable outside the jurisdiction of the person represented in the deepfakes.⁶⁰ Due to the internet-based dissemination, there is a great chance that the victim is in a different country from the content creator⁶¹ and, since there is no specific international crime standard, legal difficulties in cross-border proceedings may be immense.

Therefore, even in a context in which it is possible to determine who carried out the violation, it can be said that the lack of a specific law and jurisdictional issues might lead to a lack of accountability of the agent and to the absence of proper reparation to the victim, leaving the pornographic deepfake victims helpless when seeking justice and assistance.⁶² Thus, relevant international human rights standards on privacy and women's rights remain unimplemented.

However, it should be noted that all the previous considerations revolve around (the difficulty of) legal actions taken after violations are made and have caused great damage, but not to prevent or mitigate them. Even if there was legislation that could be applied and the person responsible punished and able to pay for the proper amount for which it was condemned, such legal prosecution would not remedy the consequences caused by videos that have gone viral.⁶³ Hence, it is important to remember that prevention is an equally important aspect related to the implementation of human rights.

An important form of prevention should consist in measures related to the platforms that create and disseminate harmful content. For example, legislation should be created to hold platforms responsible for allowing the harmful use of deepfakes

⁵⁸ Gabriel (n 54) 866; Emily Poole, 'Fighting Back Against Non-Consensual Pornography' (2015) 49 University of San Francisco Law Review 181.

⁵⁹ Mathilde Pavis, 'Rebalancing Our Regulatory Response to Deepfakes with Performers' Rights' (2021) 27(4) *Convergence: The International Journal of Research into New Media Technologies* 974.

⁶⁰ Johanna Gibson, 'Where Have You Been? CGI Film Stars and Reanimation Horrors' (2020) 10(1) Queen Mary Journal of Intellectual Property 1.

⁶¹ Henry, Flynn, and Powell (n 36) 571.

⁶² Sarah Esther Lageson, Suzy McElrath, and Krissinda Ellen Palmer, 'Gendered Public Support for Criminalizing "Revenge Porn"' (2019) 14(5) *Feminist Criminology* 560.

⁶³ Edvinas Meskys and others, 'Regulating Deep Fakes: Legal and Ethical Considerations' (2020) 15(1) *Journal of Intellectual Property Law and Practice* 24.

for pornographic purposes. Currently, some companies rule out content review through their internal regulations,⁶⁴ but the state's prevention duties would force the state to overrule that policy. The current lack of regulation allows for the prioritisation of commercial interests over ethical interests and social justice goals.⁶⁵

The protection of these rights by digital platforms, it may be contended, should not exist only if there is explicit legislation that obliges them to do so. Although there is currently no binding treaty that obliges the private sector to commit to human rights, corporations are advised to respect those rights willingly in accordance with the United Nations Guiding Principles on Business and Human Rights (UNGPs).⁶⁶

The participation of corporations in the defence of human rights is increasingly present in debates on Internet governance.⁶⁷ There is an understanding that internet intermediaries have a responsibility to their immediate stakeholders and to Internet users who may be directly or indirectly affected by their practices.⁶⁸ Under these logics, it is worthy to mention the recommendation for internet intermediaries made by the Special Rapporteur on Violence against Women calling for the protection and elimination of discrimination and violence against them in the digital space.⁶⁹

Indeed, some platforms on which deepfake practices occur have already taken some measures aiming to tackle issues related to these problematic uses. As this technology's software spreads and improves, the difficulty to create convincing videos decreases.⁷⁰ Both Twitter and Reddit have hence pledged to banish the distribution of pornographic images of celebrities generated using AI.⁷¹ Reddit even updated its rules on involuntary pornography to include the ban on 'representations that have been faked', in addition to prohibiting the publication of images

⁶⁴ As in the case of the Pornhub platform, 'Terms Of Service' (*Pornhub*, 25 April 2022) <<https://pt.pornhub.com/information/terms#terms>>.

⁶⁵ Henry and Witt (n 15) 761.

⁶⁶ Human Rights Council, 'Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie—Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework' (21 March 2011) UN Doc A/HRC/17/31.

⁶⁷ Emily Laidlaw, 'Myth or Promise? The Corporate Social Responsibilities of Online Service Providers for Human Rights' in Mariarosaria Taddeo and Luciano Floridi (eds), *The Responsibilities of Online Service Providers* (Springer 2017) 135.

⁶⁸ Rikke Frank Jørgensen and Anja Møller Pedersen, 'Online Service Providers as Human Rights Arbitrers' in Mariarosaria Taddeo and Luciano Floridi (eds), *The Responsibilities of Online Service Providers* (Springer 2017) 31.

⁶⁹ Office of the United Nations High Commissioner for Human Rights (OHCHR), 'Report of the Special Rapporteur on Violence Against Women, Its Causes and Consequences on Online Violence Against Women and Girls From a Human Rights Perspective' (18 June 2018) UN Doc A/HRC/38/47.

⁷⁰ John LaMonaga, 'A Break From Reality: Modernizing Authentication Standards for Digital Video Evidence in the Era of Deepfakes' (2020) 69(6) American University Washington College of Law 1945.

⁷¹ Janko Roettgers, 'Reddit, Twitter Ban Deepfake Celebrity Porn Videos' (*Nasdaq*, 7 February 2018) <www.nasdaq.com/articles/reddit-twitter-ban-deepfake-celebrity-porn-videos-2018-02-07>.

'with the specific purpose of falsifying explicit content' or soliciting 'lookalike pornography'.⁷²

However, some studies suggest that real changes in corporate policy often require a public scandal that companies cannot ignore.⁷³ Social pressure alone may not work,⁷⁴ which reinforces the need for a proactive regulatory framework in this regard.

Effective ways to combat the creation of this content could derive from banning the publication of this type of content altogether. Enforcement of such a regulation is feasible since tech companies usually have the technical knowledge and resources to create deepfake detection technologies.⁷⁵ These technologies could alternatively be developed in a format that exposes their creation model, that is, with an open source, thus ensuring greater transparency, privacy protection standards, and allowing a better way to deal with this problem. Such regulation could be derived both from the companies' own desire to comply with human rights and from state duties to effectively deal with human rights violations.

Another solution that may be taken by the platforms themselves, especially with government subsidies and support, could be to develop specific tools to report this type of video.⁷⁶ Although some websites offer the possibility of reporting fake pornographic content, it is noteworthy that it is currently the victim's responsibility to report each of the irregular URLs and provide evidence that the image is false and was disseminated in a non-consensual manner (while, in contrast, these same search engines typically allow the upload and access of content titled 'fake porn' without any filter or limitation).

Another mitigation strategy to combat misuse may lie again in a change made by the platforms themselves in their systems. Although it is difficult to imagine a law with this provision, it is possible to think about the implementation of policies to mobilise companies in favour of the protection of these rights. Platforms, where deepfake is known to be distributed, could start to proactively control the content.⁷⁷ They could also, through changes in their systems, implement a filter that prevents the manipulation of images that involve nudity and pornography when these are

⁷² Reddit, 'Never Post Intimate or Sexually Explicit Media of Someone Without Their Consent' (Reddit, 7 March 2022) <www.reddithelp.com/hc/en-us/articles/360043513411>.

⁷³ ARTICLE 19, 'Public Interest, Private Infrastructure' (Article 19, 2018) <www.article19.org/resources/public-interest-private-infrastructure/>.

⁷⁴ Indeed, companies are learning how to deal with scandals linked to human rights violations on their platforms without engaging in meaningful changes in their way of functioning, as mentioned in Mike Ananny and Tarleton Gillespie, 'Public Platforms: Beyond the Cycle of Shocks and Exceptions' (The Internet, Policy and Politics Conference, Oxford University, 9 July 2016) <<http://ipp.ox.ac.uk/sites/ipp/files/documents/anannyGillespie-publicPlatforms-oiii-submittedSept8.pdf>>; Giles Moss and Heather Ford, 'How Accountable Are Digital Platforms?' in William H Dutton (ed), *A Research Agenda for Digital Politics* (Edward Elgar 2020) 97.

⁷⁵ Meskys and others (n 63) 12.

⁷⁶ Henry, Flynn, and Powell (n 36) 578.

⁷⁷ See the chapter by Giovanni De Gregorio and Pietro Dunn in this volume.

recognised by their system. Although normally platforms only moderate the content after the material is published, and do not proactively examine the content posted on them,⁷⁸ some exceptions should be considered. Some platforms have indeed developed automated systems that can detect and remove harmful online content, including deepfake content, before anyone has access to or can share the material.⁷⁹ In that same vein, all platforms should find and remove image-based sexual abuse content,⁸⁰ preventing further viewing and/or sharing of these images.

Changing the default option with respect to fake porn material by way of making the consent by the featured person obligatory is also a possibility. Platforms could furthermore use their database to verify that the video has not already been reported as harmful, preventing its republication. These measures, if implemented progressively or in combination, could lead to a better fight against gender-based violence in a privacy-violating scope.

Although some of these options seem laborious, together they could prevent the proliferation of fake porn victims, as well as psychological or professional consequences for many women. Thus, companies and platforms that contribute to the production or dissemination of deepfake can and should mitigate the misuse of this type of AI. A dialogue between public authorities and digital platforms to combat the misuse of deepfake, plus legislation that regulates the performance of platforms in dealing with these abuses, could demonstrate not only the commitment of states in the implementation of human rights but also the human rights commitment of these platforms.

5 Conclusion

As discussed, technological advances allow the film and entertainment industry to advance rapidly. These novelties, however, are not restricted to content creators at the cinematographic level but are also marketed as products offered to a much larger audience. Deepfake, for instance, is marketed through different applications, for different purposes, on multiple platforms.

⁷⁸ Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 Harvard Law Review 1599; Sarah T Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale UP 2019).

⁷⁹ This, is in fact, what Facebook does, having a tool, developed in 2019, that was designed to detect non-consensual intimate images, remove them, and present some sort of support to victims. It is AI that aims to detect 'almost nude' images shared non-consensually. The AI is trained to recognise patterns of language and words, updating a database of previously confirmed non-consensual intimate images. Also, a comparison is made with complaints already received by the platform. In the event of being recognised as a harmful product, the user is notified and may have their account suspended. As mentioned in Antigone Davis, 'Detecting Non-Consensual Intimate Images and Supporting Victims' (*Meta*, 15 March 2019) <<https://about.fb.com/news/2019/03/detecting-non-consensual-intimate-images/>>.

⁸⁰ Henry and Witt (n 15) 759.

The popularisation of this technology has a negative consequence: the creation of pornographic deepfakes, which occurs without the consent or knowledge of the victims, exposing them to a situation of vulnerability and exploitation. This type of image manipulation goes beyond the limit of mere unpleasantness, representing a form of gender discrimination and violation of the human right to privacy.

Advanced technology allows for a level of editing that makes the image look real, causing serious harm to victims. Even when the editing does not reach high accuracy the mere association of a woman's face in a context that leaves her vulnerable or brings her shame is enough to cause damage.

States are obliged to protect their population from violations of privacy and gender discrimination. However, seeing as these rights have not been properly implemented in protecting against the indiscriminate use of deepfake, joint human rights action against companies and platforms that allow and disseminate harmful content produced could represent a form of human rights protection. In this regard, both those who create and administer these apps can limit the use of deepfake in contexts where the technology is used for harmful purposes. Necessary limitations may hence arise either from national regulations or from the companies' own desire to align themselves with human rights principles, even where they face no binding duty to do so.

The deepfake technology is not an enemy in itself. Its role within the context of entertainment enables a playful online environment and allows the creation of otherwise impossible audience experiences. However, the creation, marketing, and distribution of this technology must be done carefully, considering its harmful potential.

Artificial Intelligence and Disability Rights

*Antonella Zarra, Silvia Favalli, and Matilde Ceron**

1 Introduction

The increasing pervasiveness of artificial intelligence (AI) raises significant questions for persons with disabilities (PWD). On the one side, AI represents an unprecedented opportunity, potentially contributing to the advancement of their rights to equality and non-discrimination, employment, access to goods and services, independent living, communication, and education. Indeed, AI solutions can act as a powerful assistive technology for PWD.¹ For instance, AI-enabled systems might enhance personal mobility and independence through navigation tools. Adaptive learning platforms might provide personalised learning experiences for students with disabilities. Eye-tracking and voice-recognition software technology might enable PWD to communicate. Robots and other tools powered by AI might provide care and other assistance within the house.² On the other hand, AI poses new challenges and problems for the protection of PWD, who are at risk of being further marginalised. AI may fail to consider the needs of PWD multidimensionally. For instance, AI solutions built according to strong ableist assumptions might eventually harm PWD, aiming to ‘fix’ something that does not reflect canonical standards. Furthermore, the nuanced nature of the concept of disability may prove problematic for the development of automated decision-making models that work according to fixed parameters. Finally, PWD are at increased risk of algorithmic bias, being excluded or discriminated against and unlikely to be represented in training data sets.

* On the whole, this article is the product of joint reflection. However, section 1 was written by Matilde Ceron, section 2 was written by Antonella Zarra, and section 3 was written by Silvia Favalli. Section 4 was written by Antonella Zarra and Matilde Ceron, while section 5 was written by Matilde Ceron, Antonella Zarra, and Silvia Favalli together.

¹ In broad terms, the World Health Organization (WHO) defines assistive technology as ‘an umbrella term covering the systems and services related to the delivery of assistive products and services’ that ‘enables people to live healthy, productive, independent, and dignified lives, and to participate in education, the labour market and civic life’. See <www.who.int/news-room/fact-sheets/detail/assistive-technology>. More specifically, in this context, we refer to any device, equipment, software program, or product that is employed to increase, maintain, or improve the functional capabilities of PWD.

² European Disability Forum, ‘Plug and Pray? A Disability Perspective on Artificial Intelligence, Automated Decision-Making and Emerging Technologies’ (2018).

Against this background, the broader risks of AI for citizens' rights and society at large have caused mounting pressure for the regulation of algorithms. In this context, vulnerable groups such as PWD and their perspectives have been chronically underrepresented in the decision-making processes. Yet, disability has gained saliency within the literature on digital inclusion, showing an ambivalent relation between technology and the rights of vulnerable groups.³ As a result, how to ensure that a disability human rights approach makes its way into the priorities of AI lawmakers is a central and timely concern, recognised also by policymakers. From such a perspective, the chapter considers how AI may positively or negatively affect the human rights of PWD enshrined in the United Nations Convention on the Rights of Persons with Disabilities (CRPD). It brings together the extant literature on the digital inclusion of people with disabilities and key AI use cases to analyse how the CRPD in its four dimensions of inclusive equality applies to AI. The analysis offers a benchmark against which to evaluate existing and proposed regulatory exercises and propose the requisites for a disability human rights-based approach to AI.

We proceed, in section 2, by identifying how AI affects the rights of PWD positively by providing assistive functionalities and negatively by exacerbating bias and discrimination. In section 3, we present the relevant legal framework of the rights of PWD within the CRPD in relation to AI to address its opportunities and threats for the redistributive, recognition, participative, and accommodating dimensions of inclusive equality. Against this benchmark, section 4 proposes a disability human rights-based approach to AI on which basis we evaluate contemporary and proposed regulatory solutions and their shortcomings. The chapter concludes by highlighting the primary challenges for the digital inclusion of PWD and proposals for an AI regulatory ecosystem better equipped for not leaving vulnerable groups behind.

2 Artificial Intelligence and Persons with Disabilities

In this chapter, we refer to PWD to identify the constituency group of those who have any type of disability. In doing so, we adopt the 'person first' terminology in accordance with the CRPD. According to the Center for Disease Control and Prevention (CDC), disability is any condition of the body or mind that makes it more difficult for someone to do certain activities and interact with the world around them.⁴ Disabilities may affect, among other things, people's vision,

³ G Goggin, 'Disability and Digital Inequalities: Rethinking Digital Divides with Disability Theory' in M Ragnedda and G Muschert (eds), *Theorizing Digital Divides* (Routledge 2018) 63–74; G Goggin, K Ellis, and W Hawkins, 'Disability at the Centre of Digital Inclusion: Assessing a New Moment in Technology and Rights' (2019) 5(3) *Communication Research and Practice* 290–303.

⁴ Center for Disease Control and Prevention (CDC), 'Disability and Health Overview' (CDC, 15 September 2020) <www.cdc.gov/ncbddd/disabilityandhealth/disability.html>.

movement, thinking, memory, learning, communication, hearing, or mental health. The World Health Organization (WHO) distinguishes three dimensions of disability.⁵ First, impairment in a person's body structure or function, or mental functioning (eg loss of memory, loss of vision). Second, activity limitations (eg difficulty seeing, hearing, walking, and problem-solving). Third, participation restrictions in normal daily activities, such as working, engaging in social and recreational activities, and obtaining health care.

Beyond any attempt of categorisation, it must be stressed that disability is not a monolithic concept. First, there is not an agreed upon definition—nor in the blackletter law nor in the scholarship—of what constitutes a ‘disability’, and often such term is pitted against the concept of ‘normalcy’ or ‘ability’. Disability studies scholars distinguish between medical and social models of disability. While the medical model, by relying on biomedical standards of normalcy, has been claimed to reinforce the stigmatisation toward PWD, who fall outside of these standards, the social model characterises disability as the result of disabling environments and attitudes, thus linking the notion of ability to social and material contexts.⁶ In this context, disability can be considered as a synonym of ‘deficiency’ and ‘dependency’ and it intersects with other characteristics that typically define marginalised groups, such as gender, ethnicity, economic status, and age. Such meaning fails to consider that people with similar disabilities may have different economic and social backgrounds, ethnicities, genders, and ages as well. In this vein, article 1 of the CRPD, ‘recognizing that disability is an evolving concept’,⁷ does not provide a precise definition of disability. It adopts an open definition (or non-definition)⁸ affirming that ‘persons with disabilities include those who have long-term physical, mental, intellectual, or sensory impairments which in interaction with various barriers may hinder their full and effective participation in society on an equal basis with others’.

It is precisely the fluid nature of disability that gives rise to many of the problems associated with the interaction between PWD and emerging technologies based on AI. In fact, many of the AI-based technologies that are either specifically or tangentially developed to assist PWD are built according to pre-defined and rigid logic rules that do not capture the nuances of the notion of disability and, in addition, they might be built based on ableist assumptions. Given that technology has the sharp potential to impact the life of PWD in many ways, it is worth assessing the role of AI-enabled tools in supporting PWD, and the challenges that emerge for them as these technologies permeate our society. Indeed, AI can improve the life of PWD, but it can also perpetuate bias and discriminatory behaviours, exacerbating

⁵ *ibid.*

⁶ T Siebers, *Disability Theory* (University of Michigan Press 2008).

⁷ CRPD, Preamble, para (e).

⁸ G De Burca, ‘The EU in the Negotiation of the UN Disability Convention’ (2010) European Law Review 174.

their marginalisation. On the one hand, AI systems may act as assistive technologies, supporting and helping PWD engage in different tasks, eventually helping them achieve a more independent living experience. On the other hand, the proliferation of automated decision-making systems equipped with the ability to infer people's features through, for instance, profiling, entails significant risks of discrimination that may dilute the positive effects of the adoption of smart devices.

2.1 The Assistive Function of AI-based Technologies

Undoubtedly, AI systems can be an emancipatory tool for PWD. They can advance equality in many fields, including employment, education, housing, and access to services and products, leading to a more independent life. Whether based on standalone software or embedded in hardware, some AI systems may work as assistive technologies improving independence in PWD. AI enables accessibility in many fields, from urban mobility to education. Some cities are leveraging AI and data to improve accessibility to sidewalks and reduce the barriers to mobility.⁹ Furthermore, smart home devices are equipped with accessibility features that help blind individuals identify objects.¹⁰ Similarly, specific AI-powered mobile applications such as *Be My Eyes*, put in contact sighted volunteers with blind and low-vision people by letting volunteers solve small tasks.¹¹ The same community can use AI-powered navigation tools, and screen reading devices. AI is also beneficial to tackle problems of communication and learning. Some solutions involve eye-tracking or speech recognition that helps PWD to better communicate. Interaction is further facilitated through speech-to-text algorithms that support people who are non-speaking. In the same vein, thanks to adaptive learning platforms, students with disabilities can modulate their learning experience according to their needs.

In addition to this, owing to its key and growing deployment in medical research, AI may also contribute to improving the diagnosis of conditions, enhancing the development of drugs as well as supporting PWD's physical rehabilitation. The most advanced algorithms developed in the healthcare sector can allow for the early diagnosis of diseases as well as for the identification of accurate treatments.¹² Assistive robots are being used to take care of people with mental health or physical conditions, while chatbot therapists are programmed to provide emotional

⁹ 'AI for Inclusive Urban Sidewalks Project' (*Smart Cities for All*, 2019) <<https://smartcities4all.org/ai-for-inclusive-sidewalks/>>. See also the chapter by Sofia Ranchordás in this volume.

¹⁰ 'These Smart Home Devices Can Enhance Independence for People with Disabilities and Mobility Needs' (*The New York Times* (29 April 2022) <www.nytimes.com/wirecutter/reviews/best-assistive-smart-home-technology-for-disabled/>).

¹¹ See <www.bemyeyes.com>.

¹² T Bergmann and others, 'Developing a Diagnostic Algorithm for the Music-Based Scale for Autism Diagnostics (MUSAD) Assessing Adults with Intellectual Disability' (2019) 49 *Journal of Autism and Developmental Disorders* 3732.

support to patients.¹³ The adoption of wearable robotic exoskeletons, for instance, enables individuals with limited mobility to walk or stand upright. Similarly, smart speakers such as Amazon's Echo that can respond to voice commands and allow individuals to switch on lights from remote, although designed for everyone, constitute a powerful enabler for a more accessible life.

2.2 The Dark Side of AI for Persons with Disabilities

While, on the one hand, technological advancements bring about substantial advantages in terms of accessibility, inclusiveness, and equality, on the other hand, they can turn out particularly risky for PWD. When assessing the drawbacks of AI-based technologies for this community, a line should be drawn between their negative consequences for already marginalised categories and the additional difficulty derived from the fluid nature of the concept of disability. In other words, PWD may be negatively affected by AI as a collateral effect of a system that—along with other vulnerable subjects—does not take them into account, or they can be further damaged as the technology is built according to rigid categorisations or ableist assumptions.

One of the main problems relates to the potential discriminatory outcomes of the decisions made by automated decision-making (ADM) systems which may be trained on data sets that do not include information on vulnerable categories such as PWD or may use fixed parameters to determine what constitutes a disability and what does not. So far, much of the literature on AI bias and fairness has been focusing on the effects on gender and race, neglecting other attributes, including disability.¹⁴ However, there is increasing evidence showing that AI systems incorporate systematic biases from social attitudes in which disability is classified as 'bad'. Machine learning (ML) algorithms programmed to detect online hate speech and make sentiment analysis classified texts mentioning disability as more toxic/negative.¹⁵ In general, it can be argued that AI bias is an umbrella problem that includes issues related to privacy, consent, and misclassification.

Concerning privacy, PWD are often reluctant to share their personal information about their condition, because they fear being discriminated against. Such an

¹³ TL Chen and others, 'Robots for Humanity: Using Assistive Robotics to Empower People with Disabilities' (2013) 20 IEEE Robotics & Automation Magazine 30 <<http://ieeexplore.ieee.org/document/6476704/>>; and Falguni Patel and others, 'Combating Depression in Students Using an Intelligent ChatBot: A Cognitive Behavioral Therapy' in 2019 IEEE 16th India Council International Conference (INDICON) (IEEE 2019) <<https://ieeexplore.ieee.org/document/9030346/>>.

¹⁴ CL Bennett and O Keyes, 'What Is the Point of Fairness? Disability, AI and the Complexity of Justice' (2020) ACM SIGACCESS Accessibility and Computing 125; JR Foulds and others, 'An Intersectional Definition of Fairness' in 2020 IEEE 36th International Conference on Data Engineering (ICDE) (IEEE 2020) <<https://ieeexplore.ieee.org/document/9101635/>>; and Shari Trewin, 'AI Fairness for People with Disabilities: Point of View' (2018) arXiv:1811.10670 <<https://arxiv.org/abs/1811.10670>>.

¹⁵ Trewin (n 14).

issue is particularly relevant in the case of algorithms employed by companies to hire people that might exacerbate discriminatory behaviours. When artificial intelligent agents are used in the hiring process, they may fail to consider the working experiences of PWD. Applicants with disabilities were systematically assigned a negative weight in the scoring process used by an AI tool to determine employability, resulting in the placement agency allotting fewer resources and less support to an applicant with a disability in his or her job search.¹⁶ These systems profile people based on, among other things, any healthcare needs. Similarly, AI is used frequently by public administrations to determine the eligibility for social protection benefits. Using biased data sets (which would fail to represent PWD or other marginalised categories) might result in PWD being potentially excluded from obtaining such services. In fact, automated decision-making systems' calculations do not allow for a case-by-case assessment, which is key in the evaluation of PWD's needs and costs of care. A similar outcome could be observed in the case of faulty automated decision-making systems employed to attribute access to private insurance. PWD might not only have such insurances denied, but also face higher premiums.

Technologies in theory designed to help and assist PWD, on the contrary, can make it more difficult for them to use smart devices. For instance, people with speech impairment are not understood by smart assistants—voice-controlled artificial agents that can engage with individuals by answering questions, playing audio, and controlling smart home devices. These smart assistants are trained on data sets that bias the understanding toward the average speech. In the same vein, in the case of autonomous vehicles, equipped with computer vision technologies, individuals in wheelchairs are frequently in danger of not being identified, neither by the system nor by the human trainer, thus being potentially run over. It must be stressed, however, that some of the most advanced technologies developed to help PWD entail strong ableist assumptions that might be counterproductive. An example of this is represented by virtual reality (VR) games equipped with gaze detection and face recognition to train people with an autism spectrum disorder to have eye contact and to respond with appropriate levels of emotions. The real audience of such technology is the medical professional entrusted with training their behaviour to fit into the mainstream idea of 'normalcy'.

Once these categories of risks have been identified, the biggest challenge stems from embedding the concept of disability in the models. Disability is an umbrella term encompassing multiple categories and subject to diverse classifications. It is not straightforward how to insert the categories of disability in the training data

¹⁶ M Buyl and others, 'Tackling Algorithmic Disability Discrimination in the Hiring Process: An Ethical, Legal and Technical Analysis' in 2022 ACM Conference on Fairness, Accountability, and Transparency (ACM 2022) <<https://dl.acm.org/doi/10.1145/3531146.3533169>>; D Allhutter and others, 'Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective' (2020) 3 *Frontiers in Big Data* 5.

sets because the notion of disability and its categories are themselves difficult to disentangle. The concept includes a variety of physical and mental health conditions that can emerge at any time in an individual's lifetime.¹⁷

When envisioning remedies and solutions to the phenomenon of algorithmic discrimination at large, experts tend to rely on the concept of AI fairness. However, ensuring AI fairness in relation to disability differs greatly from ensuring AI fairness for other protected attributes such as gender, ethnicity, or age. The biggest dissimilarity is the nuanced way in which disability can manifest. Furthermore, disability information is sensitive and not easily shared, because of fears of discrimination.¹⁸ Moreover, some disability activists call for abandoning the concept of fairness to embrace the one of justice.¹⁹ According to them, fairness frameworks focus on equality and presume universality, thus failing to address the structural oppression faced by PWD. On the contrary, justice presumes power systems, acknowledging the inherent inequality that comes with equitable opportunity and centres on the most marginalised.²⁰

A final concerning aspect originates from biometric recognition tools, where PWD are often overlooked in the design process. For instance, voice recognition systems might fail to identify the voice of people with speaking issues, or emotion recognition systems, trained to detect anger, fear, or sadness with biased data may wrongly identify traits of PWD.

3 Artificial Intelligence and the Convention on the Rights of Persons with Disabilities

In the last few years, several United Nations specialised agencies are drawing attention to the impact of AI on human rights. Most notably, the Office of the United Nations High Commissioner for Human Rights (OHCHR),²¹ the United Nations Children's Fund (UNICEF),²² the International Labour Organization (ILO),²³ and the United Nations Educational, Scientific and Cultural Organization (UNESCO)²⁴ analysed the balance of risks and opportunities presented by AI, also mentioning

¹⁷ Trewin (n 14) 1.

¹⁸ ibid.

¹⁹ AL Hoffmann, 'Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse' (2019) 22 *Information, Communication & Society* 900.

²⁰ Bennett and Keyes (n 14) 1.

²¹ OHCHR, 'The Right to Privacy in the Digital Age. Report of the United Nations High Commissioner for Human Rights' UN Doc A/HRC/48/31 (13 September 2021).

²² UC Berkeley Human Rights Center Research Team and UNICEF, 'Memorandum on Artificial Intelligence and Child Rights' (30 April 2019).

²³ International Labour Organization (ILO), 'World Employment and Social Outlook 2021: The Role of Digital Labour Platforms in Transforming the World of Work' (Geneva 2021).

²⁴ United Nations Educational, Scientific and Cultural Organization (UNESCO), 'Report of the Social and Human Sciences Commission (SHS)' UN Doc 41 C/73 (22 November 2021).

the disproportionate risks faced by PWD. In the same vein, the UN Committee on Economic, Social and Cultural Rights' General Comment No 25 (2020) on science and economic, social, and cultural rights,²⁵ recognised that 'persons with disabilities have suffered deep discrimination in the enjoyment of the right to participate in and to enjoy the benefits of scientific progress and its applications' and recalls the necessity to 'bring their unique perspectives and experiences into the scientific landscape'.

Moreover, at the regional level, several initiatives have started considering AI and human rights in the Economic and Social Commission for Asia and the Pacific²⁶ and in the African Commission on Human and Peoples' Rights.²⁷ In Europe, the Council of Europe is drafting a Convention on AI, human rights, democracy and the rule of law,²⁸ while the European Union (EU) has published a proposal for an AI Act to regulate the use of AI.²⁹ Nonetheless, as the Special Rapporteur on the rights of PWD points out, to date, 'there has been little detailed assessment of the direct benefits and potential harms of artificial intelligence for the world's approximately 1 billion persons with disabilities'.³⁰

The CRPD represents the legal benchmark against which to assess the risks and opportunities of AI. This is the first UN human rights convention recognising PWD as autonomous human rights holders, so affirming the human rights dimension of disability rights.³¹ The CRPD provides a complete enumeration of all the human rights of PWD, from which it is possible to infer the legal obligations on states on the development and use of AI. Moreover, it provides a robust responsibility of states parties to incentivise and regulate the private sector in eliminating discrimination on the basis of disability 'by any person, organization or private enterprise'.³²

²⁵ United Nations Committee on Economic, Social and Cultural Rights (CESCR), 'General comment No 25 on Science and Economic, Social and Cultural Rights (Art 15(1)(b), (2), (3) and (4) of the International Covenant on Economic, Social and Cultural Rights)' UN Doc E/C.12/GC/25 (30 April 2020), para 34.

²⁶ Economic and Social Commission for Asia and the Pacific (ESCAP), 'Artificial Intelligence in Asia and the Pacific' (November 2017).

²⁷ African Commission on Human and Peoples' Rights (ACHPR), '473 Resolution on the Need to Undertake a Study on Human and Peoples' Rights and Artificial Intelligence (AI), Robotics and Other New and Emerging Technologies in Africa', ACHPR/Res. 473 EXT.OS/XXXI (2021).

²⁸ Council of Europe Committee on Artificial Intelligence (CAI), 'Revised Zero Draft [Framework] Convention On Artificial Intelligence, Human Rights, Democracy And The Rule Of Law' CAI(2023)01 <<https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193d>>.

²⁹ European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>

³⁰ Human Rights Council, 'Report of the Special Rapporteur on the Rights of Persons with Disabilities' UN Doc A/HRC/49/52 (28 December 2021), 6.

³¹ See eg R Cera, V Della Fina, and G Palmissano, *The United Nations Convention on the Rights of Persons with Disabilities: A Commentary* (Springer 2017); I Bantekas, MA Stein, and D Anastasious, *The UN Convention on the Rights of Persons with Disabilities: A Commentary* (OUP 2018).

³² CRPD, art 4(1)(e).

The main concerns refer to the unequal treatment or discrimination based on disability deriving from the use of AI. Algorithmic profiling reproduces and amplifies intersectional forms of discrimination, also contributing to the creation of new patterns of discrimination.³³ With reference to disability discrimination, as argued above, AI models often exclude PWD due to the failure in addressing disability directly in the development of original data sets and models,³⁴ in adopting accommodating solutions in the collection and analysis of proxy data, or due to digital accessibility issues.³⁵ Equality and non-discrimination are at the heart of the CRPD and permeate the entire convention. States parties are required to promote equality and prohibit all forms of discrimination on the ground of disability (article 5), while reasonable accommodations must be provided to enable an individual to fully exercise their rights (article 2). States are also required to ensure the accessibility of, *inter alia*, information and communication technologies (ICT) and new technologies, as a precondition to achieve equality for users with disabilities (article 9). Furthermore, to guarantee the equalisation of opportunities (article 3), all rights in the CRPD are to be secured on an equal basis with others. Overall, the CRPD embraces a new substantive model of ‘inclusive equality’, which elaborates on four different dimensions of equality: redistributive, recognition, participative, and accommodating.³⁶

First of all, the ‘fair redistributive dimension’ of equality aims at addressing the socio-economic disadvantages PWD often must face in society. In this line, the CRPD recognises to PWD the right to an adequate standard of living and social protection, and to the continuous improvement of living conditions, without discrimination based on disability (article 28). However, AI-enabled technologies used to automatically ascertain individuals having the right to access government-funded services poses PWD at risk to be illegitimately excluded from the selection when data sets are not modelled accordingly.³⁷

Similar risks are posed in the use of AI in health care, where such technology can be used, for instance, for individualising patient treatment recommendations and determining access to health insurance.³⁸ However, patients with disabilities might experience discrimination in health care (articles 25 and 26), where AI systems

³³ R Xenidis, ‘Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience’ (2021) 27 Maastricht Journal of European and Comparative Law 736.

³⁴ K Nakamura, ‘My Algorithms Have Determined You’re Not Human: AI-ML, Reverse Turing-Tests, and the Disability Experience’ in *21st International ACM SIGACCESS Conference on Computers and Accessibility* (2019).

³⁵ Human Rights Council (n 30) 15.

³⁶ Committee on the Rights of Persons with Disabilities, ‘General comment No 6 on Equality and Non-Discrimination’ CRPD/C/GC/6 (28 April 2018), para 11.

³⁷ NG Packin, ‘Disability Discrimination Using Artificial Intelligence Systems and Social Scoring: Can We Disable Digital Bias?’ (2021) 8 Journal of International and Comparative Law 487.

³⁸ A Vasudeva, NA Sheikh, and S Sahu, ‘International Classification of Functioning, Disability, and Health Augmented by Telemedicine and Artificial Intelligence for Assessment of Functional Disability’ (2021) 10(10) Journal of Family Medicine and Primary Care 3535.

may be programmed to reach outcomes, such as cost-cutting, which can be indirectly detrimental to PWD.

Second, the 'recognition dimension' of equality refers to the construction of the identity of PWD. It purports to combat stigma and promote the recognition of their inherent dignity and intersectionality. The CRPD affirms the right to recognition before the law (article 12) and the right to obtain, possess, and utilise documentation of their nationality or other documentation of identification (article 18). However, the use of biometric technology to facilitate legal proof of identity might constitute a barrier for PWD, where such technologies are not accessible and alternative means are not provided.³⁹

Other issues might arise concerning privacy and data protection (article 22) of data belonging to PWD as part of their digital identity and their collection for statistics (article 31). In the context of AI, users with disabilities must be adequately supported in managing their personal data and be able to maintain agency over them.

Furthermore, the 'participative dimension' of equality promotes full inclusion in the society of PWD. For this purpose, the CRPD promotes the right to freedom of expression and opinion and access to information (article 21), which is a precondition to the enjoyment of the right to participate in political and public life (article 29). States are required to take appropriate measures to provide information in accessible formats, as well as to urge private entities and the mass media to promote the accessibility of information. Such a purpose can be facilitated using AI-enabled tools, which allow for wider accessibility. Moreover, online voting systems enabled by AI technologies might be a powerful instrument in guaranteeing voting rights for PWD. However, if not designed taking accessibility in mind, they might become a further barrier for voters with disabilities.⁴⁰

Finally, the 'accommodating dimension' of equality refers to the necessity to consider the individual's circumstances when identifying tailored solutions to make space for diversity as a matter of human dignity. Accordingly, reasonable accommodations must be provided to workers with disabilities to guarantee their right to work (article 27). Such a duty must be performed starting from the recruiting process, which is increasingly implemented through the use of AI technologies such as video screening and resume-mining tools. This means that candidates with disabilities risk being excluded from the selection process based on atypical attributes even before meeting a human interviewer. Similar issues might arise about the growing use of AI-enabled worker-management platforms.⁴¹ In compliance with the CRPD, employers are required to use AI tools to avoid the potentially

³⁹ J Lazar and MA Stein, *Disability, Human Rights, and Information Technology* (University of Pennsylvania Press 2017) 204.

⁴⁰ Human Rights Council (n 30) 13.

⁴¹ M Whittaker and others, 'Disability, Bias, and AI' (AI Now Institute 2019) <<https://ainowinstitute.org/publication/disabilitybiasai-2019>>.

discriminatory impact of inaccessible technologies (article 9) and to adopt adequate reasonable accommodations (article 2) to not restrict workers with disabilities to show their skills. Similarly, reasonable accommodations must be adopted in using AI systems in education, to develop individualised support for learners with disabilities and fully achieve the right to inclusive education (article 24).⁴²

4 New Regulatory Solutions on AI and Their Impacts on Disability

The international regulatory landscape around AI is quite fragmented and heterogeneous. Most of the national regimes rely on soft law strategies with no legal obligations. In these legal frameworks, currently, there are no legal safeguards directly addressing PWD. In a report, the UN High Commissioner for Human Rights included PWD in the categories that must be particularly considered when carrying out human rights due diligence to assess any disproportionate effects when AI systems are deployed by states and businesses.⁴³ The report called for the adoption of human rights impact assessment throughout the life cycle of the systems as well as for a moratorium on the use of high-risk systems until the threats to human rights can be better mitigated.⁴⁴ While several policy initiatives on AI have been launched in the majority of countries globally, the EU was the first one to attempt to regulate holistically AI-based technologies through the Proposal for an AI Regulation (the AI Act), presented by the European Commission in 2021. Although the Regulation is not yet in force and will be subject to changes in the final stages of negotiations, it constitutes the first attempt of binding legislation on AI ever drafted; thus, it is worth analysing it through the lenses of disability rights.⁴⁵

The regulation adopts a layered risk-based approach, classifying AI into categories and imposing obligations on AI systems based on their level of potential or intrinsic harm. More specifically, the AI Act prohibits harmful AI practices that are a threat to people's safety and rights because of the unacceptable risk they create. PWD are explicitly mentioned in article 5, which bans those AI systems that, among other things, exploit vulnerable groups with mental or physical disabilities. While the Act does not require any accessibility obligations for the placing in the market of AI systems, Title IX (article 69) encourages businesses to adopt codes of conduct in which voluntary accessibility commitments would be

⁴² R Kohli and others, 'Artificial Intelligence Technology to Help Students with Disabilities: Promises and Implications for Teaching and Learning' in A Singh and others (eds), *Handbook of Research on Critical Issues in Special Education for School Rehabilitation Practices* (IGI Global 2021) 238–55.

⁴³ OHCHR (n 21) para 49. See also the chapter by Isabel Ebert and Lisa Hsin in this volume.

⁴⁴ *ibid.*

⁴⁵ At the time of writing, the final text of the Regulation has not been approved yet, therefore our analysis is based on the Commission's proposal adopted on 21 April 2021.

included. The same title calls for ‘stakeholders’ participation in the design and development of the AI systems and diversity of development teams’. Overall, however, while acknowledging the risks posed by AI to PWD in the context of employment and access to certain essential private and public services, the proposal falls short of ensuring that PWD are directly involved or taken into consideration when developing AI solutions. At the same time, some limitations of AI systems that would not be resolved by including PWD remain, such as their rigid classification of the parameters that define what is to be considered ‘disability’.

The consultation on the AI Act, ahead of the tabling of the Commission proposal, offers an additional benchmark of the concerns of interest groups of PWD in relation to AI and the proposed regulation. In this context, the European Union of the Deaf, the European Blind Union and the European Disability Forum submitted responses and position papers to the consultation, highlighting aligned concerns in relation to AI and rights of PWD.⁴⁶ The key concerns refer to accessibility and non-discrimination, with scepticism toward approaches relying on voluntary measures and self-regulation. Specifically, human rights impact assessments and oversight from independent bodies and civil society are deemed fundamental to mitigate the risk of business-centred approaches, calling, for example, for ex ante conformity assessment under the presumption of exclusionary and discriminatory risks associated with AI. Conversely, the proposed sphere of technology considered high-risk may be insufficient especially in a context of widespread bias throughout the life cycle of AI, failing to include PWD in the selection of use cases, among developers and in the training data. Stakeholders confirm two further problematic arenas in relation to AI and disability, which may remain a shortcoming within the proposed regulation. Firstly, a concern emerges in the absence of a mandate for universal design which may be problematic both for accessibility and the stigma associated with being relegated to the use of ad-hoc assistive technology. Additionally, biometric data is confirmed as especially dangerous for PWD not only because of the risk of exclusion if its use is inaccessible but especially as it may allow for the disclosure of sensitive data concerning disability status and lead to discrimination. Against such a benchmark, even the most advanced regulatory attempt for human-centric approach to AI—as it stands in the proposal of the European Commission—may be insufficient to fully protect the rights of PWD.

⁴⁶ European Disability Forum, ‘EDF position on the European Commission’s White Paper on Artificial Intelligence: A European Approach to Excellence and Trust’ (2020) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/public-consultation_en>.

5 Conclusion

This chapter explored how AI offers challenges and opportunities for the full realisation of the rights of PWD. The legal system is designed to ensure the inclusion and the rights of PWD in the development of AI systems. In this context, the CRPD sets forth multiple dimensions of equality relevant to AI applications. The AI implications of the CRPD along the dimensions of redistributive, recognition, participative, and accommodating equality in turn offer a benchmark for regulatory efforts supporting the protection of the rights of PWD. A human rights-based approach to the regulation of AI would need to deliver guarantees that do not exacerbate the socio-economic disadvantages of PWD, nor hinder their identity or identification. Moreover, it would fully embrace the participation of PWD throughout the AI life cycle and would not pose a barrier for effective accommodations to be put in place in crucial areas such as employment and education.

A critical aspect surrounds the element of representation. Specifically, the lack of inclusion of PWD in all phases of the AI life cycle—from the development of the models and the data on which they are trained to the design of the policies⁴⁷—exacerbates significantly the likelihood of harm for PWD and their marginalisation. Further structural limitations include the lack of intersectionality and the lack of a unique legal definition of disability. In this context, for systems that will make or influence decisions affecting human lives, a broad range of stakeholders must be involved in their development, including PWD. Inclusive representation offers valuable guidance for developers in identifying possible implications of the technology and testing the technology's performance on edge cases and under-represented populations. In addition, broad inclusiveness allows accounting for the fluid nature of the concept of disability, which is often less about physical or mental impairments than it is about how society responds to impairments.

Additionally, the regulation of AI should prioritise human-centric goals. In this context, the target population and the reference values of developers are embedded in the system they create. As a result, from the perspective of fostering inclusiveness, a central question is the extent to which those that are affected by the deployment of the system are reflected in the value and target references of developers. Such a question is especially central in the context of vulnerable groups such as PWD whose needs may vary from the general population, may be impacted in ways differently, and/or face additional barriers. Failing to include PWD in the early phases of AI development hence risks further exclusionary impacts down the line. Indeed, the limited representation of PWD in the training and evaluation of AI systems is highly detrimental. Moreover, their inclusion opens the complementary challenge of protecting the sensitive personal data of PWDs, especially on

⁴⁷ For a comprehensive overview and explanation of the phases of the entire AI life cycle, see the chapter by Martina Šmuclerová, Luboš Král, and Jan Drchal in this volume.

the disclosure of their disability status. Yet, in the trade-offs between inclusion and protection of personal data, there is the risk to legitimise the adoption of a paternalistic approach. From such a perspective, the aspect of agency represents a key priority in enabling accessibility and safeguarding PWD from discrimination. As a result, regulatory efforts should underline the importance of users' control over the data used by AI applications, especially crucial for PWD.

To conclude, the chapter highlights how ongoing regulatory efforts represent at the same time an opportunity to promote the full inclusive equality of PWD while a potential source of further discrimination if the human rights of these vulnerable groups are not satisfactorily addressed. The report of the High Commissioner of Human Rights sets a positive outlook for the recognition of the need for a human-centric approach on par with the inclusion of PWD. However, awareness of policymakers is far from enough to deliver effective inclusion of the needs of PWD in the regulation of AI. On this account, the EU invites a far less optimistic prospect. Indeed, even in the context of a regulatory effort aimed at deploying a human rights-centric approach, outcomes may be underwhelming for PWD. It is no surprise that the absence of stringent enforcement mechanisms has been fiercely criticised by interest groups of PWD in the stakeholder consultation ahead of the European Commission Proposal for an AI Act. Moreover, technologies that would be banned under the proposed regulation are far narrower than what interest groups identify as a potential threat to PWD.

As a result, while AI can undeniably offer unprecedented opportunities for the inclusion of PWD, threats abound from a disability human rights approach. If the participation of PWD has emerged as a key requirement for the deployment of AI that supports their inclusion rather than discrimination, the same consideration extends to ongoing regulatory developments. At the policy level, failing to embed the concerns of PWD in this new regulatory wave poses a threat to the ability to equip our societies with an AI ecosystem which can support the inclusive equality of PWD and potentially other vulnerable groups.

PART V

ARTIFICIAL INTELLIGENCE AND
FAIR PROCEDURE

18

Artificial Intelligence and Fair Trial Rights

Helga Molbæk-Steenig and Alexandre Quemy

1 Introduction

Everyone is entitled in full equality to a fair and public hearing by an independent and impartial tribunal, in the determination of his rights and obligations and of any criminal charge against him.¹

Everyone has a right to a fair trial in both civil and criminal cases. The right is prescribed in the International Covenant on Civil and Political Rights (ICCPR)² as well as in regional human rights instruments such as the European Convention on Human Rights (ECHR)³ and the American Convention on Human Rights (ACHR)⁴ and in numerous regional and national constitutions and bills of rights.⁵ A cornerstone of the rule of law and a foundational prerequisite for securing many other fundamental rights, the right to a fair trial is procedural in nature, and central in the international system of human rights protection as a whole. It is also, by far, the most frequently adjudicated right and violation found before international human rights courts. Taking the European Court of Human Rights (ECtHR) as an example, as of April 2022, more than half of all cases in its online database (HUDOC)⁶ contained a complaint under article 6 and some 11,300 cases had resulted in a fair-trial related violation. By comparison, the next most frequently violated articles 5 (personal freedom) and 3 (torture, degrading, and inhuman treatment) each account for around 3,000 violations.

¹ Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR), art 10.

² International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR), art 14.

³ Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR), art 6.

⁴ Organization of American States (OAS), American Convention on Human Rights (ACHR), art 8.

⁵ Eg European Union, Charter of Fundamental Rights of the European Union (2012), art 47; United States, Constitution of the United States: The Bill of Rights (1789), Sixth Amendment; Basic Law for the Federal Republic of Germany (2014 [1949]), art 103.

⁶ European Court of Human Rights (ECtHR), HUDOC database (1960–2022) <hudoc.echr.coe.int>.

There are important differences in how court systems work in different jurisdictions. Some make use of juries and lay judges, others rely exclusively on professional judges. In some systems, the judge sits alone, in others, adjudication takes place in colleges. In some cases, the judge(s) have an investigative role while in others they decide exclusively based on the statements given by the parties. In some systems judgments include reasoning and separate opinions may be annexed, in other systems they are brief determinations of the results reached. The right to a fair trial as prescribed in international human rights conventions does not favour any one such system over another, but regardless of the system, it requires that certain foundational elements are met. These include: that the trial must be conducted by an independent and impartial tribunal established by law conducting a public hearing,⁷ within a reasonable time and without undue delay.⁸ For criminal trials, it is furthermore required that the defendant is presumed innocent until proven guilty,⁹ that they are informed promptly and in a language they understand about the accusations against them,¹⁰ and that they have adequate time and resources to mount a defence, including access to free legal assistance if they cannot pay themselves.¹¹ The accused also have a right to have witnesses heard in their favour and to an interpreter if necessary.¹² Together these rights are sometimes referred to as ‘equality of arms’.¹³ The right to a fair trial is also closely related to the rights not to be tried for the same offence twice and not to be convicted on the basis of retroactive law.¹⁴ Together, these rights coalesce into a few fundamental principles. One is temporal: trials take place after the fact based on events that took place before the trial and laws in existence before the events took place. Judgments are also individual, even if several accused can be sentenced during the same trial. Given the right to a presumption of innocence and the right to mount a defence, a person cannot be sanctioned on the basis of a judgment against another. This precludes both collective punishments and situations in which the outcome of one trial implicates guilt in another. It also implies a right of non-discrimination integral to the right to a fair trial both in the sense of consistency—like cases must be treated alike—and in the sense that trial outcomes may not be influenced by any group belongings of the parties to the case.

Unfortunately, securing the right to a fair trial is fraught with difficulties. In this chapter, we will delve into two recurring problems identified in adjudication and in academic literature, both of which have been identified as potential

⁷ ICCPR, art 14(1); ECHR, art 6(1); ACHR, art 8(1), (5).

⁸ ICCPR, art 14(1), (2)(c); ECHR, art 6(1); ACHR, art 8(1).

⁹ ICCPR, art 14(2); ECHR, art 6(2); ACHR, art 8(2).

¹⁰ ICCPR, art 14(3)(a); ECHR, art 6(2)(a); ACHR, art 8(1)(a).

¹¹ ICCPR, art 14(3)(b), (d); ECHR, art 6(2)(b), (c); ACHR, art 8(1)(c)–(e).

¹² ICCPR, art 14(3)(e), (f); ECHR, art 6(2)(d), (e); ACHR, art 8(1)(b), (f).

¹³ *Öcalan v Turkey* App no 46221/99 (ECtHR, 12 May 2005), para 140.

¹⁴ ICCPR, arts 14(7), 15; ECHR, art 7 and Protocol 7 (Protocol No 7 to the Convention for the Protection of Human Rights and Fundamental Freedoms), art 4; ACHR, arts 8(4), 9.

spaces for AI intervention and assistance: namely the substantive problem of securing consistent and impartial adjudication, and the practical difficulties with securing justice within a reasonable time. The first problem has been the subject of intense academic interest for almost as long as courts and judges have existed and raises foundational questions about the nature of justice and the limits of human cognition.¹⁵ The other, though less widely studied academically, is the most frequent human rights complaint¹⁶ and a serious threat to the international human rights system itself. As international human rights adjudication has become better integrated in more and more societies, international and regional human rights bodies are also experiencing an increase in incoming applications. The ECtHR which allows direct individual applications to the greatest degree of the existing courts is the international body furthest ahead in this development and the increased workload is particularly evident here as it is currently battling a backlog of 76,700 cases, habitually taking years to deal with cases.¹⁷

In section 2, we will go through existing challenges related to securing consistent and impartial adjudication within a reasonable time and the potential of AI in alleviating them. Section 3 examines the AI applications that are already applied in various jurisdictions and the problems that have been identified with their usage. Section 4 builds a theoretical framework for distinguishing between judicial tasks that can be beneficially outsourced to an AI and tasks that cannot. This section combines knowledge about the nature of AI applications with legal theory development with a starting point in Ronald Dworkin's conceptualisation of the ideal judge 'Hercules'. Section 5 will review existing regulation in comparison with this framework.

2 Biased Human Judges and Inefficient Courts

The fact that the Court has taken eight years to deliver its judgment is in itself inexcusable ... Switzerland has amended its legislation in the meantime. In such a situation, I wonder whether it is useful to have a judgment of the Court finding a violation of the Convention.¹⁸

¹⁵ Eg D Kahneman, O Sibony, and CR Sunstein, *Noise: A Flaw in Human Judgment* (Little Brown 2021).

¹⁶ F Edel, *The Length of Civil and Criminal Proceedings in the Case-Law of the European Court of Human Rights* (Human Rights Files No 16, Council of Europe 2007).

¹⁷ As of April 2023. Statistics on pending cases according to the Department for the Execution of Judgments of the ECtHR. Updates monthly at <www.echr.coe.int/Pages/home.aspx?p=reports&c=>>.

¹⁸ Dissenting Opinion of Judge Keller, joined by Judge Popović in *Ruiz Rivera v Switzerland* App no 8300/06 (ECtHR, Judgment of 18 February 2014).

Courts of all kinds rely on legitimacy for their power to make decisions. They do not wield power themselves but rely on others to execute their judgments. In the case of international human rights courts, execution relies on the very states that the court may have just judged to have violated a human right. As does the courts' funding. The source(s) of this legitimacy is a field of discussion in and of itself, but few commentators would likely disagree with former President of the ECtHR, Robert Spano, that important elements include accountability, transparency, equality, and non-discrimination.¹⁹ The ability to consistently render impartial judgments based on the law and within a reasonable time is almost too foundational to merit consideration. Both are nonetheless imperfectly implemented in many if not most judicial systems. The quote above from the dissenting opinion by Judge Keller in *Ruiz Rivera v Switzerland* illustrates one problem with lengthy court cases; they can render the judgment useless once it arrives. 'Justice delayed is justice denied' is a popular expression of the importance of speedy trials.²⁰ Another problem is that being involved in court cases is resource intensive and stressful for litigants and it can hamper the full enjoyment of many other human rights including the freedom of movement, the freedom of speech, and the right to property.

With regards to securing impartiality and non-discrimination, scholars within the school of legal realism have uncovered problems in this regard as well. For example, Michæl Benesty and Anthony Sypniewski uncovered massive differences between the rates at which individual judges in the French administrative courts of appeal granted asylum,²¹ a finding so unpopular that France enacted a law banning the use of statistics and machine learning (ML) in the study of individual judges.²² Similar studies have been carried out in the United States (US) with troubling results.²³ In addition to uncovering variance, scholars have investigated which factors other than the law might influence decision-making and have found correlations between the tendency of leniency or strictness and the weather, the time of day, and whether a local sports team had recently won or lost.²⁴ When the goal is to secure a

¹⁹ Rosalind English and Robert Spano, *New Strasbourg Court President on AI and the law* (UK Human Rights Blog, 22 May 2020) <<https://ukhumanrightsblog.com/2020/05/22/new-strasbourg-court-president-on-ai-and-the-law/>>.

²⁰ Sometimes attributed to nineteenth-century British statesman, William Ewart Gladstone.

²¹ Michael Benesty, 'L'Open Data et l'Open Source, des Soutiens Necessaires à une Justice Predictive Fiable' (2017) 5 Journal of Open Access to Law 1–11; Michael Benesty, 'The Impartiality of Some French Judges Undermined by Machine Learning' (SupraLegem (Medium), 2016) <<https://medium.com/@supralegem/the-impartiality-of-some-judges-undermined-by-artificial-intelligence-c54cac85c4c4>>.

²² Malcolm Langford and Mikael Rask Madsen, 'France Criminalises Research on Judges' (*Verfassungsblog: On Matters Constitutional*, 22 June 2019) <<https://verfassungsblog.de/france-criminalises-research-on-judges/>>.

²³ Daniel L Chen and others, 'Early Predictability of Asylum Court Decisions' (March 2017) TSE Working Paper 17/781; Kahneman, Sibony, and Sunstein (n 15) 73.

²⁴ There are entire monographs on these issues, see eg Jesper Ryberg, *Domstolens blinde øje: Om betydningen af ubevidste biaser i retssystemet* (Djøf 2016); Brian M Barry, *How Judges Judge: Empirical Insights Into Judicial Decision-Making* (Routledge 2020).

fair and impartial judgment and equal treatment of all litigants, these types of noise are almost as troubling as consistent bias. ‘Noise’ is a term of art borrowed from Kahneman, Sibony, and Sunstein.²⁵ It refers to inconsistencies in decision-making that are not predictably biased in one direction or another. As such, a biased judge would be consistently stricter than other judges either towards all defendants or only towards those belonging to a specific gender, ethnic group, sexuality, or other status. Meanwhile, a noisy judge might not be consistently biased against any one group, but would be overly prone to letting their hunger, frustration about a local sports team, or other irrelevant interferences impact their overall strictness. It is important to note that inconsistencies are not always evidence of problems; deeply unfair judicial systems can be very consistent and predictable; legal realist Karl Llewellyn proposed as an example that judges might employ an unwritten but unbending rule that ‘the mining company always wins’ regardless of the legal questions sought answered.²⁶

Long before different types of noise and bias became known by these names, judicial systems have incorporated structures attempting to limit them. One such structure is the obligation to give reasons, which provides the judge with an occasion to self-check their intuitions.²⁷ Another approach for combating biases is the use of the bench, that is, having decisions made not by a single judge but by a college of judges with or without the inclusion of laypersons. This is how the Human Rights Committee, the ECtHR, and the Inter-American Court are set up. The bench may act as a forum where the intuitions of the different judges meet, as well as a catalyst for the giving of reasons. At the ECtHR, interviews with judges show that they particularly value inputs from judges from different professional backgrounds from their own, because when different intuitions meet, more thorough treatment of the case follows.²⁸ The benefit of the bench, however, can be watered down by, for example, the tendency of groups to yield to seniority.²⁹ There is also a string of research suggesting that groups tend to reach more extreme outcomes than individuals do on their own.³⁰

Inconsistencies as those described above are indicators that the applicants and defendants in question may not have received a fair trial, and they are detrimental

²⁵ Kahneman, Sibony, and Sunstein (n 15) 1–20 and 73ff on differentiating between noise and bias.

²⁶ Frederick Schauer, *Thinking Like a Lawyer* (Harvard UP 2009) 132–33.

²⁷ Frederick Schauer, ‘Giving Reasons’ (1995) 47 Stanford Law Review 633; Kahneman, Sibony, and Sunstein (n 15) 281–83.

²⁸ Fred J Bruinsma, ‘Judicial Identities in the European Court of Human Rights’ in Aukje van Hoek and others (eds), *Multilevel Governance in Enforcement and Adjudication* (Intersentia 2006) 203, 217; Kanstantsin Dzehtsiarou and Alex Schwartz, ‘Electing Team Strasbourg: Professional Diversity on the European Court of Human Rights and Why it Matters’ (2020) 21 German Law Journal 621, 628.

²⁹ Tomer Broude, ‘Behavioral International Law’ (2014) 163 University of Pennsylvania Law Review 1099, 1127–48.

³⁰ Eleanor C Main and Thomas G Walker, ‘Choice Shifts and Extreme Behavior: Judicial Review in the Federal Courts’ (1973) 91 Journal of Social Psychology 215; Kahneman, Sibony, and Sunstein (n 15) 94–104.

to court legitimacy. Consequently, there are good reasons for human rights scholars and advocates to be concerned with the inadequacies of existing judicial systems. The hope that artificial intelligence (AI) and ML applications might assist in overcoming the problem of inefficiency and bias has been around since computers first became commonplace. If the role of a judge is to deliver predictable and consistent decisions, algorithms might be able to do the job faster, and might be better at eliminating problematic human biases that have nothing to do with the law. Dystopian visions of futuristic robotic judges and pre-emptive crime fighting have,³¹ however, been around for just as long. Today, ML algorithms are employed as decision support systems (DSS) by police, judicial bodies, and administrations in numerous jurisdictions worldwide.³² In the future, they might also assist the ECtHR, which has its own problem with backlogged cases,³³ and according to some commenters with coherence and consistency as well.³⁴ Tom Zwart has even suggested that 'strays from existing case law seem to be caused by the fact that the members of the Court have simply lost track of their own case law because there are too many judgments'.³⁵

A central question remains about what DSS are in fact doing and what data they base their results on. Depending on the data input and the parameters given to the AI, algorithmic decisions may be just as biased as human ones, may miss subtle but vital clues, or may be modelled on outdated or incomplete data.³⁶ In section 3, we look into the kinds of DSS that have been applied in real-life scenarios until now and whether they have solved the problems of inconsistency, lack of impartiality, and efficiency.

³¹ Eg *Minority Report* (1956) by Philip K Dick.

³² Fabio Chiusi and others (eds), *Automating Society* (Bertelsmann Stiftung and AlgorithmWatch 2020).

³³ Mikael Rask Madsen and Robert Spano, 'Authority and Legitimacy of the European Court of Human Rights: Interview with Robert Spano, President of the European Court of Human Rights' (2020) 1(2) European Convention on Human Rights Law Review 165, 179; Helga Molbæk-Stensig, 'AI at the European Court of Human Rights: technological improvement or leaving justice by the wayside?' *Ordine Internazionale e diritti umani* 5, 1254–67.

³⁴ Lord Lester of Herne Hill, 'Universality versus Subsidiarity: A Reply' (1998) 3 European Human Rights Law Review 73, 75.

³⁵ Tom Zwart, 'More Human Rights than Court: Why the Legitimacy of the European Court of Human Rights is in Need of Repair and How It Can Be Done' in Spyridon Flogaitis, Julie Fraser, and Tom Zwart (eds), *The European Court of Human Rights and Its Discontents: Turning Criticism into Strength* (Edward Elgar 2013) 71, 86.

³⁶ Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Penguin 2016) 20–25.

3 Biased and Unjust Artificial Judges

[T]he irresistible attractions for Governments to move in this direction are acknowledged, but [there is also a] grave risk of stumbling, zombie-like, into a digital welfare dystopia.³⁷

If humans are inconsistent, biased, slow, easily distracted, and prone to letting their own convictions interfere with their interpretation of legal facts and principles, might an AI be a more appropriate mouth of the law? Some of the authors we encountered in section 2 suggest that this may be the case, though at present none suggest a complete replacement of the human judge. As an example, Jon Kleinberg and others suggest that bail decisions made by a computer could reduce detention by up to 42 per cent without increasing crime or the incident of skipping bail and that it would be more racially fair, jailing 41 per cent less people of colour;³⁸ while Daniel Chen suggests using AI to nudge judges to consider judging in ways that are more similar to their colleagues.³⁹ At the same time, there is a rich literature concerned with the way AI and ML are already being used in judicial and administrative systems, such as Cathy O’Neil’s book on the American jurisdiction which lists several problematic usages in policing and in sentencing. She points to the problem of discrimination and transparency caused by the proprietary nature of ML algorithms.⁴⁰ Another example is Jesper Ryberg who argues from a more foundational theoretical starting point that an over-reliance on predictive tools may blind authorities to questions of why states punish individuals in the first place.⁴¹ The European Union’s strategy on AI also labels the usage of AI in judicial systems among its ‘high-risk’ applications.⁴²

An illustrative case displaying these concerns is *State v Loomis* which was decided by the Wisconsin Supreme Court in 2016.⁴³ The case was launched when a defendant in a drive-by shooting case, Eric Loomis, filed a complaint after initial sentencing in the Circuit Court, claiming that the Court’s reliance on the predictive

³⁷ United Nations General Assembly (UNGA), Report of the Special Rapporteur on Extreme Poverty and Human Rights: The Digital Welfare State (2019) UN Doc A/74/493.

³⁸ Jon Kleinberg and others, ‘Human Decisions and Machine Predictions’ (2018) 133 Quarterly Journal of Economics 273, 1–15, 3.

³⁹ Daniel L Chen, ‘Incremental AI’ (2022) American Journal of Evaluation.

⁴⁰ O’Neil (n 36).

⁴¹ Jesper Ryberg, ‘Risk-Based Sentencing and Predictive Accuracy’ (2020) 23 Ethical Theory and Moral Practice 271.

⁴² European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts’ (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD), Annex II, para 8(a) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

⁴³ *State v Loomis* 2015AP157-CR (Wisconsin Supreme Court, 2016); Editorial, ‘*State v Loomis*: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing’ (2017) 130 Harvard Law Review 1530.

sentencing software COMPAS was a violation of his right to due process. He argued that his right to individualised sentencing had been violated and that the inclusion of gender in COMPAS' risk assessment violated his right to non-discrimination. Loomis lost the case. While the Wisconsin Supreme Court acknowledged that COMPAS assesses group data, it still maintained that the practitioners guide to the software had warned the users of COMPAS that it predicted group behaviour, not individualised behaviour.⁴⁴ It also stated that in any case enough discretion had been left to the Circuit Court, which was aware of the limitations of COMPAS,⁴⁵ and consequently the right to individual sentencing had not been violated.⁴⁶ On the topic of gender, the proprietary nature of COMPAS became particularly evident. Since COMPAS keeps its algorithm a business secret, the state and defendant disagreed about whether the algorithm actually factored in gender, and if so, exactly how.⁴⁷

Although Loomis ultimately lost the case, it illustrates potential problems with the use of such software quite well. First and foremost, the case demonstrates the specific problem of secrecy in the use of algorithms and secondly, the more general problem of knowledge gaps among legal professionals using ML software to conduct their work. A more recent European case on the use of algorithms in the tracking of social fraud, shows, however, that even when the data input for the algorithm are not secret as such, the problem of transparency may remain. In the *SyRi* case from the Netherlands, an ML algorithm was used on all receivers of welfare benefits to pick out cases for fraud investigation. Following an application from a privacy-focused NGO (NJCM), the Hague Court determined that the use of private data in the *SyRi* software on this massive scale was a violation of article 8 of the ECHR.⁴⁸ The Court also noted that the state had deliberately failed to explain clearly how the software conducted its risk assessment and drew connections on the basis of the input data, making it impossible for the Court to fully understand what it was dealing with.⁴⁹ The Hague Court thus showed an important sense of humility that the Wisconsin Court lacked; recognising that the ability to critically understand ML algorithms rests with a small, educated elite—and a different one than the one usually engaged in handing out judgments, leading to the Wisconsin Circuit Court repurposing a recidivism risk software for initial sentencing without even attempting to recalibrate or recode.

A more general problem with the current usage of ML in judicial systems is that of 'objective ignorance'—that is, the problem of the future. Even though algorithms may provide reliable predictions on average or group behaviour, they still get the

⁴⁴ *State v Loomis* (n 43) para 69.

⁴⁵ *ibid* para 109.

⁴⁶ *ibid* paras 120–21.

⁴⁷ *ibid* para 76.

⁴⁸ *SyRi* case C-09-550982-HA ZA 18-388 (Rechtbank Den Haag, 2020).

⁴⁹ *ibid* paras 6.46, 6.49.

future wrong in complex systems, real-life (non-toy), and individual cases more often than not.⁵⁰ One problem is causation. While a computer can generally spot a correlation between temperature increase during the summer and a rise in ice cream sales much faster than a human being can, it cannot tell whether the temperature rose because of the increase in sales or the other way around. In simple situations like this, human beings often have no problems.⁵¹ The problem arises in more complex situations, as well as situations where bias may interfere. It might not be immediately clear to human decision-makers whether a given geographical area has a higher rate of arrests because it has more crime, or if it is because that area is more heavily policed.⁵²

Another problem is that legal systems are dynamic. New rules and new precedents can erase or diminish the power of old rules and precedents. Today's ML algorithms learn by incorporating as many examples as possible of a stationary object or process, but they cannot unlearn. A pet-recognising AI that has been fed pictures of cats and dogs will get progressively better at recognising these animals as pets, but it would have to be retrained from scratch if there was suddenly a rule change that only cats should be recognised as pets. Meanwhile, a human child can unlearn recognising dogs as friendly pets immediately after having been bitten. Furthermore, the retrained AI would have to exclude the previous pieces of knowledge about four-legged fluffiness as if they did not exist, with all the problems that implies, far less data, far less accuracy, and no leveraging or 'connecting the dots' between the past decisions, now invalid, and the new ones.

4 Creating Hercules: A Framework for the Division of Labour Between Human and Computer

[I]t seems wrong to imprison a man awaiting trial on the basis of a prediction that he might commit further crimes if released on bail. For any such prediction, if it is sound, must be based on the view that an individual is a member of a class having particular features, which is more likely than others to commit crime ... But it is unjust to put someone in jail on the basis of a judgment about a class, however accurate, because that denies his claim to equal respect as an individual.⁵³

⁵⁰ Kahneman, Sibony, and Sunstein (n 15) 143–44.

⁵¹ Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Penguin 2018) 4–10.

⁵² O'Neil (n 36) 84–104.

⁵³ Ronald Dworkin, *Taking Rights Seriously* (first published 1977, Harvard UP [1981]) 13.

A central question for anyone wishing to automate any part of the work of the judge is what it is in fact that a judge does, including whether he or she is finding or making the law. In other words, is there (usually or always)⁵⁴ a ‘correct’ legal answer to every question? If there is, it is the task of the judge to find that answer, which in turn is a process that may lend itself to varying degrees of ‘automation’ or ‘assistance’ by algorithm. If, however, there are legal vacuums where no law is applicable, judges are left with a law-making task. It is in the nature of courts that they must always reach a conclusion, even when they are in doubt. It is already contested whether the court can legitimately take on such a law-making task, and it would unquestionably be a democratic problem outsourcing law-making to an automated system. For the purposes of this discussion, we will therefore focus on legal theories that work with an initial assumption that law is found not made. There are various schools of thought on how exactly the law is or ought to be found when clear legal guidance runs out; from originalists with a preference for preparatory documents and textualists concerned with specific wording, to followers of teleological strategies letting interpretation be guided by the object and purpose of the law. In this chapter, however, we will focus on the potential of the coherence-focused rules- and principles-based ‘right answer theory’ provided by Ronald Dworkin. There are several reasons for this. One is that the theory is generally applicable in national and international, civil, and criminal law, as well as human rights law (HRL). Another reason is its explicit engagement with the failings and insufficiencies of human judges when pursuing the right legal answer, providing an obvious opening for a place for discussing whether an artificial judge might be of help.

A particularly useful element in Dworkin’s theoretical framework for the study of the potential of AI is a method for delimitating between policies and laws. This is helpful because it provides for delimitating what the work we are aiming to automate is and what it is not, and it relates to foundational questions about the legitimacy of courts. Dworkin’s right answer thesis is first and foremost prescriptive, claiming not that every legal question *is* answered correctly, but that every question theoretically has one and only one correct answer. Since human beings are fallible and often fail to discover this right answer, he proposed a thought experiment utilising a super-human judge by the name of Hercules.⁵⁵ Hercules has infinite knowledge of the written law and all historical precedent. If a similar case has been decided before, he would therefore never make the mistake of missing a vertically binding or *stare decisis* precedent. It should go without saying that Hercules is completely unbiased, and his judgments are unaffected by weather, local sports teams, or what he had for breakfast. Many of the features that make Hercules the

⁵⁴ *ibid* 81; Ronald Dworkin, ‘Judicial Discretion’ (1963) 60 *The Journal of Philosophy* 624; HLA Hart, *The Concept of Law* (first published 1961, Clarendon [1994]) 272.

⁵⁵ Dworkin (n 53) 105–30.

ideal judge could be found in modern AI systems; unparalleled ability to process immense data including all precedents and laws within a particular jurisdiction, but they are not the only features of Hercules.

Dworkin states that Hercules works with a ‘right answer’ thesis in mind with which he can expand interpretations in hard cases where there is no clear guidance from statutory law and binding precedent. This is done by generalising principles already found in the law, but not by creating new policies. The reason for this is that the law and the principles it contains generate rights and duties for individuals, whereas policies are societal compromises aiming at striking out a direction for society, which is in the realm of democratic politics. Dworkin’s theory lists several important differentiating factors between policies and principles including a temporal one; that is, policies make predictions about the future while the application of the law is oriented backwards in time.⁵⁶ Another important difference is that policies are collectively minded, seeking desirable goals based on predictions about what a group on average will do in specific situations, while at the court the individual is measured as responsible solely for their own actions in relation to general rules.⁵⁷ When deciding policy, legislators, therefore, make more or less educated guesses about how society will react to a given change in the law. They might think that crime will become less prevalent if punishments were increased, or that lower taxes will increase the number of hours people are willing to work. But because societies are complex and because the future is uncertain, these policies might work, or they might fail. It may turn out that lower taxes instead incentivise people to work less while having the same standard of living, or that longer prison sentences make it harder for former criminals to re-enter society as law-abiding citizens. It makes sense, Dworkin argues, to have politicians make these sorts of predictions because they necessarily contain priorities and policy choices which the electorate ought to be able to reject or accept at regular intervals.⁵⁸ Courts on the other hand derive their legitimacy by declaring what *is* not what *might be*, and therefore it is reasonable for judges to determine their application without interference from an electorate.

Some of the ML applications currently in use have mixed up this divide between the legal and political realm. The recidivism-predicting software COMPAS and similar applications provide policy-style predictions on potential recidivism, and they do this on the basis of group-predictors which takes away from the determination of individual rights and responsibilities which is fundamental to the nature of law.⁵⁹ This is already problematic when the predictions are used for bail decisions, but it is particularly worrisome when used in the initial sentencing and when it

⁵⁶ *ibid* 85–90.

⁵⁷ *ibid* 12–15.

⁵⁸ *ibid* 84–88.

⁵⁹ *ibid* 12–13.

incorporates facts about the defendant and their environment which are not legally relevant and would be inadmissible in a regular court.⁶⁰ Hercules is purposefully not envisioned with an ability to see the future beyond the aim of securing consistency in the case law because any party to a court case has a right to a particular judgment (the right answer) based on existing law and the facts of the case.

Another problem that emerges when squaring the use of AIs for decision-making with the task of Hercules is that computers by design can only apply things that are clearly defined, but many of the principles that Hercules may use to deal with hard cases are by nature vague and applied to a degree rather than either/or. If the legislator had not predicted that a specific type of case could emerge, the ML application cannot be prepared for it either. Furthermore, one of the justice system's purposes is giving offenders an opportunity to reform and rehabilitate. This is turned on its head when the risk assessment is based on group-based factors the individual cannot change him/herself.

One place where an ML application could help human judges become more like Hercules is in the application of some tasks performed by legal practitioners which require abilities for which the machine is already capable of doing better in terms of volume, such as reading and remembering more documents and estimating probability distribution. ML applications of this kind are sometimes known as 'cognitive computing' as they extend the ability of the human brain.⁶¹ Hercules' infinite knowledge of the written law and historical case law would make many human lawyers and judges jealous, but digitalisation and simple searches have already improved their access to this knowledge, and more sophisticated computing may improve it even further. Another problem that legal scholars and practitioners face, is which precedents to rely on. Dworkin spends a few pages discussing how Hercules would handle a precedent decided by an imperfect judge who might not have found the right answer, but in many cases, scholars and practitioners are also faced with a problem of an overwhelming amount of case law, some of which might point in one direction while other point in another. Here ML applications may help make sense of case law and categorise it in accordance with the articles in question, the types of questions engaged with, or indeed whether case law is moving in one direction or another.⁶² Other tasks could entail the use of ML in automatic anonymisation of judgments for immediate publication in case-law databases, the usage of citation networks for discovering the most important case law,⁶³ or

⁶⁰ O'Neil (n 36).

⁶¹ Peter Sommer, 'Artificial Intelligence, Machine Learning and Cognitive Computing' (*IBM Blog*, 20 November 2017) <www.ibm.com/blogs/nordic-msp/artificial-intelligence-machine-learning-cognitive-computing/>.

⁶² The work undertaken by the ECHR-OD project is an example of such an approach. See Alexandre Quemy and Robert Wrembel, 'On Integrating and Classifying Legal Text Documents' in *Database and Expert Systems Applications* (Conference Proceedings 2020) 385.

⁶³ Urška Šadl and Henrik Palmer Olsen, 'Can Quantitative Methods Complement Doctrinal Legal Studies? Using Citation Network and Corpus Linguistic Analysis to Understand International Courts' (2017) 30 *Leiden Journal of International Law* 327.

assistance for first drafting of premises sections both with regards to facts and as to the law. It should not include predictions based on group behaviour. The problem with such predictions cannot be resolved by improving the algorithms, because the main issue is not the accuracy of programs such as COMPAS for initial sentencing, the problem is that the service it offers is not what courts are for. For actors other than the court, there are of course other options, the framework presented here does not exclude, for example, the use of predictive software on the part of lawyers for determining whether a civil case is worth pursuing.⁶⁴

5 Attempts to Regulate

In some countries the development of databases is alleged to have helped change the legal reasoning of practitioners, whose argumentation is less principle-based and more case-based as a result of the profusion of references to past judgements.⁶⁵

Emerging legislation, from soft-law guidelines and recommendations to potential European Union legislation, attempt to strike a balance between the potential efficiency and consistency gains from using simple and complex computing in the field of justice, and the risks to judicial independence and incorporation of systemic biases. Doing this, proposed legislation often incorporates to some extent the principles suggested by the framework above. The Council of Europe's Commission for the Efficiency of Justice (CEPEJ) has been promoting the use of information technologies generally in the administration of justice since 2016, but initially delegated more focus to the uptake of simple infrastructure technology such as file-sharing and communication via court websites.⁶⁶ On the topic of the use of simple DSS, such as templates and databases, this initial work raised concerns that there was a risk that the technology, through the ordering of case-law results or through autofill suggestions might surreptitiously impact the independence of the judge.⁶⁷ The solution—suggested in the 2016 guidelines and developed further in the 2018 European Ethical Charter on the use of Artificial Intelligence in Judicial Systems and their Environment—included an input and an output side in five general principles. Namely ensuring legal-ethical control with the initial technical input, that is, securing that the development of the DSS and AI is done with a clear understanding of what it is that the court does or is supposed to do, with a

⁶⁴ Chiusi and others (n 32) 150.

⁶⁵ Council of Europe—European Commission for the Efficiency of Justice (CEPEJ), *Guidelines On How To Drive Change Towards Cyberjustice: Stock-Taking of Tools Deployed and Summary of Good Practices* (Council of Europe 2016).

⁶⁶ *ibid* paras 1–41.

⁶⁷ *ibid* paras 44, 47–51.

focus on securing human rights and non-discrimination.⁶⁸ On the output side of the equation was instead user control—including user education allowing judges and clerks to understand what the DSS does and what it does not do as well as transparency of the technical details of the applications in use including periodical external auditing.⁶⁹ The Ethical Charter also included a principle that the data used for model input and training had to be certified and complete.⁷⁰ An example of what such transparency might look like can be found in the ECtHR's Open Data project,⁷¹ notably via the complete documentation of its process (both for the data and algorithm part), the public availability of input, output, and intermediate files, and the open-source nature of the software.

As such the principles in the CEPEJ Charter through their simplicity cover the needs for AI development for the judicial system quite well. This simplicity also results in a lack of details however, and the Charter is missing both the legal-philosophical clarity of what a fair and impartial tribunal would look like with and without the help of AI, as well as the technical specifics necessary to implement the five principles. One problem, which is evident in the report that accompanied the Charter, is that while the principles aim to be generally applicable to any legal system, there is a danger when formulating such rules to unduly and without reflection favour one particular judicial tradition over others. In the report, the knowledge foundation is very clearly based in the French system. This is evident in overarching claims that do not resonate in other systems, such as the notion that requiring judges to give reasons when deviating from the general trend in the case law would be 'tantamount to removing them from office'.⁷² Such a stance on the practice of giving reasons would be contentious in common law- and mixed systems, without ML or AI having even entered the discussion.

In 2021, the CEPEJ took initial steps towards starting a pilot project to create an assessment tool for human rights compliance of ML applications intended for use in the judicial system as well as an advisory board charged with keeping a register of existing applications and resources directed towards training and certification of users of AI systems in the judicial context.⁷³ The CEPEJ follow-up in 2021 referenced the European Union proposal for general regulation of AI. The proposed EU legislation indicates that 'AI systems intended to assist judicial authorities in

⁶⁸ Council of Europe—European Commission for the Efficiency of Justice (CEPEJ), *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment* (Council of Europe 2018) arts 1, 2.

⁶⁹ ibid arts 4, 5.

⁷⁰ ibid art 3.

⁷¹ Quemey and Wrembel (n 62).

⁷² X Ronsin and others, *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment: Appendix I—In-Depth Study on the Use of AI in Judicial Systems, Notably AI Applications Processing Judicial Decisions and Data* (CEPEJ 2018) para 35d.

⁷³ Council of Europe—European Commission for the Efficiency of Justice (CEPEJ), *Revised Roadmap for Ensuring an Appropriate Follow-Up of the CEPEJ Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment* (Council of Europe 2021).

researching and interpreting facts and the law and in applying the law to a concrete set of facts' should be considered high risk and therefore subject to increased oversight, but are not prohibited.⁷⁴ In terms of oversight, the EU legislation envisions a combination of:

- i. explicated risk assessments and management (article 9),
- ii. requirements of representativeness and completeness of training, validation, and testing data (article 10),
- iii. and requirements to provide technical documentation and record keeping (articles 11 and 12) before and during the usage of the application.

The EU legislation also requires transparency with regards to the purpose and to the technical details with regards to accuracy and robustness for users of the software (article 13). It is not clear whether article 13 would outlaw the repurposing of applications from one intended purpose (such as assessing the risk of a person skipping bail) to another purpose (such as aiding in determining the length of a sentence) as was done in *State v Loomis*.

6 Conclusion

The results achieved by AIs are in reality unrelated to the question of the legal conformity of a particular solution and cannot discriminate between legal and illegal arguments.⁷⁵

There is ample evidence that our justice systems do not function perfectly. Academic literature has uncovered suspicious correlations between judicial outcomes and judicially irrelevant facts, and fair trial rights are the most frequently violated human right in international and regional systems.⁷⁶ There may therefore be great potential for the use of AI and other ML applications in the administration of justice to provide fair trials within a reasonable time, but there are also risks involved if algorithms are trained uncritically on existing practice. What this chapter has aimed to demonstrate is that in order to unlock the potential of AI for the securing of the right to a fair trial, a clear understanding of the aim of the justice system is foundational. Decision support systems will inevitably guide legal practitioners, especially when they, as is the case in nearly all judicial systems, are short of time and resources. This has the potential to increase consistency and coherence

⁷⁴ EU Artificial Intelligence Act 2022 (n 42) Explanatory Memorandum, para 40.

⁷⁵ Ronsin and others (n 72) para 80.

⁷⁶ See introduction and statistics on pending cases according to the Department for the Execution of Judgments of the ECtHR above.

in the case law which is generally assumed to be beneficial for securing equality before the law, but without a clear understanding of what the correct administration of justice looks like, the resulting consistency could end up relying unreasonably on one parameter over others, or it could be influenced by the same biases that influence human judges, only more effectively and consistently.

In this chapter, we suggested the use of Ronald Dworkin's right answer thesis as a model for creating a human-AI hybrid that more closely resembles the perfect judge Hercules than humans alone are capable of. Dworkin developed his theory in the American common law jurisdiction, and as a result, his Hercules has a strong preference for making decisions that contribute to the general coherence of the case law. Other jurisdictions may have different priorities in the determination of justice, which was evident in the report accompanying the CEPEJ Ethical Charter.⁷⁷ Regardless of the priorities of any one jurisdiction, the solution is to develop a clear model of what the correct administration of justice looks like, since without that clarity, any attempts to incorporate AI will be fraught with risks and will not work as intended. We also suggested that there is a clear distinction between applications of ML to the decision-making process and ML used to enhance the capabilities of judges. While the former displays the current limitations of ML techniques common to many fields such as the explainability, bias, and non-stationary environment, the latter, cognitive computing, offers many opportunities to positively impact the justice domain. We also noted that emerging legislation appears to aim to regulate guided by similar concerns as those raised by our framework; however, a challenge remains in translating general legal principles into governance checklists, that is to say, to ensure that the tools used by the different actors of the law are suitable to achieve their aims, both with regards to technical soundness, the quality and completeness of the data gathered and processed, and the minimal human rights-based guarantees expected from such tools.

⁷⁷ ibid para 35.

19

Artificial Intelligence and Data Analytics

A Recipe for Human Rights Violations

Migle Laukyte

1 Introduction

One central concern in relation to artificial intelligence (AI) is the use of machine learning (ML) algorithms capable of processing massive amounts of data to extract useful information from it. These algorithms have already become part of the lawyer's toolkit, and one use that draws particular attention, for those who work with and care about human rights, is that of court analytics.

Court analytics uses dedicated algorithms to analyse different aspects of judicial proceedings within a particular jurisdiction. For instance, these algorithms can analyse the data on cases,¹ on a party's legal history in the justice system, or—and this is the focus of this chapter—on judge analytics, which detect patterns in a particular judge's rulings, the arguments he or she is most receptive to, the language he or she uses, and suchlike. All this information enables these systems to predict the possible outcome of cases, compare different judges, mitigate the risk of litigation, and devise a winning strategy.

From a human rights perspective,² judge analytics in particular, and court analytics in general, raise a lot of concerns, and here I address some of them. If, on the one hand, the analysis of court decisions makes justice more transparent and accessible,³ on the other hand, it is not clear how it affects judges, their impartiality, autonomy, and procedural rights of citizens.

There is a broader concern involved as well, namely, how to cope with a scenario in which court analytics are extended to other domains, such as healthcare, education, and others: how might this affect the individual's privacy and data protection or the right to work? One's autonomy in taking decisions? If, for instance, analytics

¹ Nikolaos Aletras and others, 'Predicting Judicial Decisions of the European Court of Human Rights: a Natural Language Processing Perspective' (2016) 2 *PeerJ Computer Science* 1 <<https://peerj.com/articles/cs-93/>>.

² The term 'human rights' is used as an inclusive term in this chapter and does not refer exclusively to the rights enshrined in the Universal Declaration of Human Rights (UDHR). See Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR).

³ Masha Medvedeva, Michel Vols, and Martijn Wieling, 'Using Machine Learning to Predict Decisions of the European Court of Human Rights' (2020) 28 *Artificial Intelligence and Law* 237.

are applied to a particular surgeon, making it possible to extract detailed metrics about her success rate in surgery, the time she spends with every patient in the recovery process, and other aspects of her practice, could—or should—this data be communicated to the patient before she agrees to a surgery? Furthermore, this data is also very appealing to the employer of the surgeon: perhaps the data about her performance could show novel ways not only to better assess the professional activities of a particular surgeon or discover bias or differentiated treatment on the basis of sex, gender, or ethnical origin of the patients, but also see novel ways to use this data in managing the hospital itself. For instance, such management could be improved by offering permanent job positions only to the most efficient (quantity of surgeries per month, less time spent with patients, and so on) surgeons versus those who prioritise not the quantity but the quality of physician–patient relationship and therefore dedicate more time to patients before and after the surgery so as to make sure that the patients are well taken care of, feel safe, and important.

The problem then is that, on the one hand, analytics can provide us with the information that may be critically important in fighting discriminatory practices that abound in our societies, but, on the other hand, analytics could also expose people to scrutiny that might undermine their autonomy, privacy, and dignity, label them and provoke negative public attitudes towards them, and also prejudice their professional future. Therefore, we have to be particularly careful in addressing this dichotomy and balancing the different interests involved: data analytics point out that the debate is no longer about an individual or a particular social group, whose rights have to be protected, but about the clash of the rights of different individuals and (professional) groups who might find themselves in the middle of the rights’ war where nobody wins and everyone loses.

Having this in mind, the chapter is organised as follows: in section 2, I address the use of AI-driven data analytics in the public domain by focusing on court and judge analytics as an emerging phenomenon related to novel uses of data flows. I address the reasons why these analytics are troublesome in terms of privacy and personal data protection, with a short reference to the European Union (EU) and, in particular, to the General Data Protection Regulation (GDPR) as an example of globally relevant legal instrument.⁴ However, I will not merely focus on the EU legal framework but consider wider implications, because the issues that data analytics raise are global and not regional: public urge for transparency, openness, trust in the national justice systems, and support for innovation are indeed international

⁴ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (General Data Protection Regulation/GDPR). On the international impact of the GDPR, see Christian Peukert and others, ‘Regulatory Spillovers and Data Governance: Evidence from the GDPR’ (2022) 41 Marketing Science 318. In this sense, any specific reference to the GDPR should be read not as a representation of regional focus on privacy and data protection, but as a global trend on how privacy and data protection could be interpreted.

claims that every society demands of its government. From there, I move on to argue that the justice system is not the only one that could be eventually affected: in section 3, I focus on the private domain and, in particular, on healthcare and argue that similar data analytics could be transferred and applied to healthcare professionals exposing them to similar, yet not identical, threats. These critical aspects that technological and AI-empowered innovation is presenting us with feed the old human rights' battles, such as finding the right balance between a variety of interests and rights when it comes to the decision-making about whose rights should prevail and why. However, I argue that data analytics highlight another—already known, yet this time framed different—aspect of power imbalance: if we open the door for data analytics without any constraints, public and private employees will be subject to continuous scrutiny not only from their patients and citizens, but also from their employers or the state itself. All these employers could take decisions affecting their employees' (professional) lives based on these analytics, which might be erroneous, showing correlations but not causations, and be in other ways misleading, misinterpreted, or taken out of context. Therefore, that data analytics is not necessarily an asset, but could easily turn into a liability for the rights of many.

2 Court and Judge Analytics

To understand what we mean when we refer to data analytics in general and the variety of its applications in specific domains, such as justice and healthcare, in particular, a definition of data analytics is in order: data analytics is a science that analyses enormous quantities and varieties of data so as to extract knowledge out of this data, and AI is a useful tool for that, as data quantities are impossible for human beings to process and therefore powerful algorithms come into play.⁵ What interests us among the variety of applications that data and AI could be put to is court analytics and, in particular, judge analytics.⁶ So how do court and judge analytics work?

Analysis of court decisions is nothing new: for instance, the decisions of the United States (US) Supreme Court have been subject to it for years because of the researchers' interest both in law clerks' role in writing them and in optimisation of judicial workload, but also in understanding how the justices vote.⁷ But simply

⁵ João Moreira, Andre Carvalho, and Tomás Horvath, *A General Introduction to Data Analytics* (Wiley 2019). For more on the relationship between data analytics and AI, see Jay Liebowitz (ed), *Data Analytics and AI* (Routledge 2020).

⁶ There are also alternative names to these analytics, such as judicial analytics, litigation analytics, and so on, that usually cover both court decisions and data about individual judges. For purposes of clarity, I will stick to the terms 'court analytics' and 'judge analytics' so as to avoid misunderstandings and use alternative terminology only when citing other authors.

⁷ Daniel Martin Katz, Michael J Bommarito II, and Josh Blackman, 'A General Approach for Predicting the Behavior of the Supreme Court of the United States' (16 January 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2463244>; Malcolm Langford, Daniel Behn, and Runar

going through the case law and using statistical methods to extract information are not the same as applying AI-based analytics: AI-based analytics as no other means enabled us not only to learn more about the courts and their case law (descriptive analytics), but also—and almost in real time—to predict court decision-making trends and performance from a continuous flow of heterogeneous data and provide solutions on how to solve identified problems (predictive and prescriptive analytics).

And this happens in Europe too: the European Court of Human Rights (ECtHR) was subject to court analytics just a few years ago, when researchers elaborated a system, based on natural language processing, to predict the outcomes of the cases.⁸ The success rate of these predictions is high, yet the most interesting aspect of the research is that court decisions—in this particular case, decisions of the ECtHR on whether there has or has not been a violation of a certain human right—are influenced by the judges present in the Chamber during the case hearing, although it is not known what he or she voted while being there.⁹ Furthermore, the research has also provided insights about how judges usually vote depending on a particular human rights violation:¹⁰ it means that a number of judges are more willing than other judges to recognise certain articles of the European Convention on Human Rights (ECHR) as more subject to violations than others.¹¹

Moving away from the US and Europe, we also discover that the Australian judicial system is addressing very similar challenges, although the Australian solution on how to make legal records publicly available causes serious difficulties to the development of judge and court analytics.¹² However, these are technical and not structural difficulties: that is to say that once technical solutions are found, Australian courts and judges risk to fall under the scrutiny of data tools as much as

Lie, 'Computational Stylometry: Predicting the Authorship of Investment Arbitration Awards' in Ryan Whalen (ed), *Computational Legal Studies: The Promise and Challenge of Data-Driven Research* (Edward Elgar 2020) 53. For an exhaustive overview of both the US and EU experiences in quantitative research of case law, see Medvedeva, Vols, and Wieling (n 3).

⁸ Aletras and others (n 1); Medvedeva, Vols, and Wieling (n 3).

⁹ Medvedeva, Vols, and Wieling (n 3).

¹⁰ ibid. It means that, for instance, some judges usually vote against the violation of the right to effective remedy (ECHR, art 13), and in favour of the violation of the freedom from discrimination (ECHR, art 14).

¹¹ We could speculate further on what it could mean. For instance, it could also prove that, for some judges, certain articles of the Conventions are violated more often than others, or which violation is easier to bring to the court, or that some of these judges are reluctant to recognise the violation of certain articles, etc. I do not want to speculate on that here, but this is the point of this chapter—the data we get from these judge analytics provides us with information that could be read in a variety of ways, some of which might be incorrect or even harmful to the credibility of the judges, since we rarely question the machines and take their suggestions, decisions, or recommendations to be the best and the most correct ones.

¹² Pamela Stewart and Anita Stuhmcke, 'Judicial Analytics and Australian Courts: A Call for National Ethical Guidelines?' (2019) 45(2) Alternative Law Journal 1.

any other judge or court in any technologically advanced and data-driven society of the world.

At this point, it might be necessary to make a distinction between court and judge analytics: in the case of court analytics, the object of study are the cases—claims, formulations, explanations, arguments, and so on—whereas, in the case of judge analytics, the object of the study is the judge as a professional. Of course, studying case law and judges might not be easily separated, yet at least theoretically the line can be drawn by focusing on *what* was decided and *how*, rather than on *who* decided the case.¹³

AI is telling us that we could learn a lot about judicial decision-making by looking at how each judge performs his or her tasks, besides positively contributing to transparency and trust of public administration in general and the judicial system in particular. Furthermore, thanks to these analytics, information about judges and their work would become easier to access for more people, thus bridging the knowledge gaps among the social groups, where more often than not justice—as many other social goods—is more accessible to those better off financially.¹⁴ McGill and Salyzyn called this emerging phenomenon ‘mainstreamed judicial analytics’.¹⁵ In fact, there are many—typically US-based—companies that offer analytical services related to the courts and judges. For instance, Thompson Reuters’ *Litigation Analytics* offers its clients such functions as: ‘Understand your judge: Get the most relevant highlights for your judge, including ruling tendencies, speed, case type experience, appeals, recent activity; Compare your judge: Quickly understand the context of your judge compared to the court average, compare judges, or apply new dynamic filters; Understand what precedents your judge has relied on in similar cases. Find out which cases your judge relies on for your issue and how often, how likely your judge is to cite to another judge from a different jurisdiction, and if there are any outliers’, and more.¹⁶

In section 2.1, I address a particular case that represents one possible way to deal with judge analytics, namely by invoking privacy and personal data protection of the judges and prohibiting judge analytics altogether. This case is one of the first examples—to the author’s knowledge—of legislative intervention aimed at stopping the judge analytics hype.

¹³ The author is aware that, even without directly addressing the work of judges, certain conclusions can still be reached or deduced from the case law that could be attributed or linked to the judge. However, it is possible to make a clear distinction between court and judge-derived analytics.

¹⁴ Michael Livermore and Dan Rockmore, ‘France Kicks Data Scientists Out of Its Court’ (*Slate*, 21 June 2019) <<https://slate.com/technology/2019/06/france-has-banned-judicial-analytics-to-analyze-the-courts.html>>.

¹⁵ Jena McGill and Amy Salyzyn, ‘Judging by Numbers: How Will Judicial Analytics Impact the Justice System and Its Stakeholders?’ (2021) 44 Dalhousie Law Journal 249.

¹⁶ This analytics is not limited to judges, but also works with attorneys and law firms. For more information, see <<https://legal.thomsonreuters.com/en/products/westlaw-edge/litigation-analytics#cite>>.

2.1 French Ban on Judge Data Analytics

As revealing and informative as the aforementioned findings may be, the judge analytics have been banned in France. Article 33 of Law 2019–222 establishes that ‘[t]he identity data of judges and members of the registry may not be reused for the purpose or effect of evaluating, analysing, comparing or predicting their actual or assumed professional practices’.¹⁷ In other words, the French legislator has put the limits on publicly available information and restricted the ways in which citizens and companies could use it.

The question is why did the French legislator prohibit these practices and why can French researchers and practitioners not freely use what until now was publicly available and open data? What seems to be a national matter in France might have serious repercussions throughout the EU and be an inspiration for other countries too, and therefore we focus on this case as it is a case that soon could turn into precedent.

French judges have already experienced what judge analytics could reveal. In 2016, Michael Benesty’s research showed that predictive algorithms, revealing information about judges, seriously question their impartiality.¹⁸ In particular, he showed that the expulsion measures called ‘Obligation to leave the French territory’ (*Obligation de quitter le territoire français* or OQTF) that oblige an asylum seeker to leave the country, were confirmed by two judges of the same court at a very different rate. In fact, one judge used this measure almost 30 per cent more often than the other. Cases in court are distributed randomly, and therefore a possible interpretation could be that regardless of the circumstances of the case, there are judges who are more eager than others to apply this measure and that the outcome of the case mainly depends on the asylum seeker’s luck not to have her case assigned to certain judges.¹⁹

This discovery seems to suggest biased judicial practices and seriously threatens the credibility of the French justice system. In this sense, analytics could be a useful diagnostic tool to assess ‘the health’ of a judicial system. For example, Rachlinski

¹⁷ The original text reads: ‘Les données d’identité des magistrats et des membres du greffe ne peuvent faire l’objet d’une réutilisation ayant pour objet ou pour effet d’évaluer, d’analyser, de comparer ou de prédire leurs pratiques professionnelles réelles ou supposées’ (art 33 of Law No 2019-222 of 23 March 2019). Violations are subject to prison sentences of up to five years.

¹⁸ Michael Benesty, ‘The Impartiality of Some French Judges Undermined by Machine Learning’ (*Medium*, 19 December 2016) <<https://medium.com/@supralegem/the-impartiality-of-some-judges-undermined-by-artificial-intelligence-c54cac85c4c4>>.

¹⁹ Similar problems in asylum adjudication in the US have been detected by Andrew I Schoenholz, Jaya Ramji-Nogales, and Philip G Schrang, ‘Refugee Roulette: Disparities in Asylum Adjudication’ (2007) 60 Stanford Law Review 295, who have applied statistical methods and discovered that whether a person will be granted asylum or not depends—among other things—also on the gender of the judge (female judges adjudicate asylum 44 per cent more than male judges). On unobserved influences on asylum courts, see Daniel L Chen, ‘Judicial Analytics and the Great Transformation of American Law’ (2019) 27 Artificial Intelligence and Law 15.

and his colleagues suggest—among other means and practices—an auditing program to check judges' discretionary determinations to not only obtain data on bias in judicial decision-making, but also to increase the accountability of judges, which otherwise is difficult to achieve.²⁰

We could only speculate what similar research could reveal if it were done in other rule-of-law abiding and democratic countries, that albeit being compliant with human rights legislation and international treaties, continue having serious issues with racism, immigration, and social inclusion. From this perspective, what Michael Benesty's research shows might be only a small yet representative sample of the reality behind immigration policies and integration programmes. This is, however, something only further judge analytics can establish.

Furthermore, looking from the perspective of data scientists, the French ban hampers innovation and raises obstacles, rather than advances the benefits of AI. In addition, analytics can reveal racial, sexual, or other kinds of bias that judges explicitly or implicitly suffer from. We already know that especially implicit racial bias is inherent in different degrees in most people and judges are no exception.²¹ The good news is that people who are aware of their biases can also take action to suppress them. This is what the researchers call 'cognitive correction', which was observed to take place especially in cases of white judges.²²

However, we should also bear in mind that things are more complicated than they might seem: to let algorithms evaluate, analyse, compare, and predict (in French law's words) the behaviour of judges can be seen in a different light as well, as discussed in section 2.2.

2.2 French Ban Through the Lens of Data Protection Laws

This prohibition does not apply to the data on court decisions as such, but it does stop data aggregation on individual judges and their performance. Therefore, court analytics—differently from judge analytics—remain legal. What the French law tries to stop is the profiling of judges, as what judge analytics does is analyse their behaviour as professionals and predict their future behaviour on the basis of this analysis, with the additional danger to base these predictions on correlations that could lead to misleading conclusions on these judges. This is what happened with potential criminal offenders and the COMPAS system in the US²³ or, even before

²⁰ Jeffrey J Rachlinski and others, 'Does Unconscious Racial Bias Affect Trial Judges?' (2009) Cornell Law Faculty Publications Paper 3/2009 786, <<https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=1691&context=facpub>>.

²¹ *ibid.*

²² *ibid.*

²³ Julia Angwin and others, 'Machine Bias' (*Propublica*, 23 May 2016) <www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

COMPAS and AI-driven analytics came into being, with the Baldus study on the death penalty,²⁴ which already showed that any data can be subject to different interpretations.

To understand the very idea of profiling and how it applies to judges, we have to look at the definitions of the GDPR. According to article 4(4) of the GDPR, profiling means ‘any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular, to analyse or predict aspects concerning that natural person’s performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements’²⁵ In the case of judges, applying data analytics means to analyse and predict their work performance as judges. In addition to that, it also means opening the possibilities to identify their ‘personal preferences or interests, reliability or behaviour’ and legal effects could be various, for example, disciplinary measures such as fines, forced transfers, or suspension.

The French legislator—not being able to ensure ‘suitable measures to safeguard’ judges’ rights, freedoms, and legitimate interests—chose to prohibit this kind of profiling altogether. From this perspective, it seems that the French legislator was the only one in the EU to take proactive action to ensure the rights and freedoms of its judges, whereas the rest of the world seems to remain passive in this regard.

Perhaps the right decision lies in a pronounced national AI strategy, which should clearly define the measures that would ensure individual rights without interfering with the goals of transparency, democracy, and innovation. The French legislator merely banned the use of analytics as concerns judges, but it could still do more by elaborating clear rules and regulations that would ensure the development and deployment of analytics based on open data of the French public administration.

However, there are two additional arguments to consider. First of all, judges are public employees and the state as an employer has a right and duty to control its employees. The research on the variety of ways this control can be carried out shows that its perception often depends on whether these evaluations are participative, learning-oriented, transparent, and enabling fair cooperation.²⁶ It is highly questionable then how judge analytics are perceived among judges and perhaps it would be correct to say that at least French judges rejected these analytics altogether. But would that be the case in other countries as well? In addition, what

²⁴ *McCleskey v Kemp* 481 US 279 (1987).

²⁵ GDPR, art 4(4). This definition is identical to the one in s 1798.140(z) of the California Civil Code. Therefore, the GDPR should be seen not as a regional legal instrument but as an international legal tool, whose spirit—rather than its exact wording—is exported to many other countries.

²⁶ Meike Wiemann, Nadine Meidert, and Antoinette Weibel, “‘Good’ and ‘Bad’ Control in Public Administration: The Impact of Performance Evaluation Systems on Employees’ Trust in the Employer” (2019) 48 *Public Personnel Management* 283.

about the individual judge's freedom to accept controls as part of their privileged role in society?

Indeed, and that takes us to the second point, although judges are public employees, they are also more than just employees because they represent one of the branches of government power, which controls and is controlled by other branches. Difficulties in maintaining the balance of powers have been discussed and debated for a long time, however, it does not mean that a viable solution to maintain the right balance has been found. As Castagnola argues, manipulation of the justice system is still a beneficial practice at least in what she calls 'new democracies' such as Argentina.²⁷ Perhaps the older democracies are not safe from this threat either inasmuch as data analytics could provide a legal tool to eliminate 'disobedient' or critical judges and promote those who row in the same direction as the government.

In section 3, I address questions about AI-based analytics in other domains, also drawing from the private sector.

3 Importing Analytics to Other Domains: Healthcare and Physician Analytics

It is not difficult to expand the considerations and doubts explained in section 2 to other domains, where the situation could get even more complicated. In this section, I address the healthcare domain where the AI-based analytics could bring forward novel aspects to already existing clashes between professionals, institutions, and citizens.

Healthcare is a particular social domain where the stakes are high and stakeholders have specific protections. For instance, physicians have very specific professional rights, such as the right to conscientious objection,²⁸ and the patients have many rights to ensure their meaningful participation in decision-making related to their health.²⁹ Nobody questions the possibilities offered by Big Data, algorithms, and other AI-based technologies in medicine,³⁰ and healthcare systems worldwide have benefited from them as much as they could afford. However, the

²⁷ Andrea Castagnola, *Manipulating Courts in New Democracies: Forcing Judges Off the Bench in Argentina* (Routledge 2018).

²⁸ There is not that much literature (to the author's knowledge) on physicians' rights as the focus nearly always falls on patients' rights. However, on physicians' rights see Ericka L Adler, 'Abusive Patient Behaviour: Physicians Have "Rights" Too' (*Physicians Practice*, 22 August 2012) <www.physicianspractice.com/view/abusive-patient-behavior-physicians-have-rights-too>.

²⁹ Besides national laws, the European Charter of Patients' Rights (2002) establishes fourteen rights of patients, <https://ec.europa.eu/health/ph_overview/co_operation/mobility/docs/health_services_c0108_en.pdf>.

³⁰ Among many, Sabyasachi Dash and others, 'Big Data in Healthcare: Management, Analysis, and Future Prospects' (2019) 6 *Journal of Big Data* 1–25. <<https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0217-0#Sec39>>.

question is not about healthcare analytics, but about how judge analytics transferred into healthcare and turned into physician analytics could affect the (already complicated) relationship between physician and patient (and her family), and—at the same time—the not less complicated relationship between the physician and the hospital she is working at.

In fact, the question is whether and to what extent such analytics should be performed. If the privacy and personal data protection aspects related to patients are all settled (data are anonymised or patients have given their consent to process their medical data and share it with third parties who develop tools for analytics), there are still further complications, similar to the ones already mentioned in section 2, but with an additional difficulty. That is to say that besides the limitations to physician profiling, there is an additional complication involving the right of the patient to be duly informed. There is an obvious clash between the rights of a physician not to be profiled and the rights of a patient to be duly informed. As a matter of fact, patients' right to information permits them 'to access to all kind of information regarding their state of health, the health services and how to use them, and all that scientific research and technological innovation makes available', as article 3 of the European Charter of Patients' Rights puts it.³¹ This is a very general description that leaves the door open to a broad interpretation of what the patients are entitled to. In fact, physician analytics is one of the many things that 'technological innovation makes available'.³²

Let us consider an example: imagine that you are waiting for a kidney surgery and get access to physician analytics regarding the surgeon who will perform the operation. You could discover a variety of metrics about her performance as a surgeon, not only the information regarding success, failure, or complication rates (regular statistical data), but also including connections that you might not have thought about, for instance, that the success rate is much higher if this surgeon is operating in the morning or that post-surgery complications emerge more often and are more fatal when the patient is female. This example raises a question as to the extent to which the patient has to be informed about the physician's performance and possible patterns or links between data that have been discovered by AI, particularly in cases such as surgeries or other treatments which could have a critical impact on patient's health and future quality of life.

However, the problem is not limited to the patient's right to be informed because even without this complicated problem, the physician could find herself under the scrutiny of its employer (hospital), which might see the need in using physician analytics as a tool to assess and evaluate the physician's performance, not necessarily

³¹ European Charter of Patients' Rights (2002) <https://ec.europa.eu/health/ph_overview/co_operation/mobility/docs/health_services_co108_en.pdf>. On the proposal to have an International Constitution of Patients' Rights, see Alison Poklaski, 'Towards an International Constitution of Patients Rights' (2016) 23 Indiana Journal of Global Legal Studies 893.

³² European Charter of Patients' Rights, art 1.

for the benefit of the physician or the patient, but for its own economic benefit. Physician analytics, based on Electronic Health Records of patients, but also data related to drug and medical material consumption by each of the employees and further data could provide the hospital with additional data on how each of the employees work, perform, and use the resources available to them.³³ Furthermore, as far as AI-based analytics are concerned, further connections between performance, results, and consumption might be established. In this regard, Adler elaborates a few proposals for the rights of physicians, and one of these proposals is ‘a right to practice medicine in a way that best evidence and experience suggests, as opposed to being forced to make decisions based on cost containment, third-party interests, or the demands of patients for particular medications, treatments tests, or referrals’.³⁴ It seems obvious then that physician analytics could lead to abuses, manipulations, and (intentional or not) misinterpretation of data.

In fact, the employer has certain powers, such as the power to guide her business in its everyday functioning, the power to adopt changes in compliance with the legal requirements related to employee rights, and also—and most importantly analytics-wise—the power to control and, if need be, sanction the employees.³⁵ Therefore, it remains to be seen how the private sector—less constrained than the public one—will use data analytics for such control. Indeed, data analytics—descriptive, prescriptive, or predictive—all offer powerful tools for control and offer the employer novel ways to exercise not necessarily legally sound authority over the employees. For instance, Spain has recently introduced changes to its Labour Code, under which firms are now required to disclose and explain the functioning of any algorithms they use in making decisions affecting their employees’ working conditions, as well as in promoting and hiring. Therefore, the question is: should we consider granting new, employment-related rights so that *datafication*—understood as the use of data analytics to anything—of one’s work performance would not prejudice him or her, or on the contrary, use any tools whatsoever if these tools could help us assess and make data- (and not prejudice or bias-) based decisions on employees, the justice system, or healthcare professionals.

4 Conclusion

I have tried to draw attention to specificities of various areas—justice and healthcare—by emphasising where the troublesome issues emerge when AI-driven data analytics are professional-oriented rather than case-oriented in both private

³³ Besides being sold to third parties, such as pharmaceutical companies.

³⁴ Adler (n 28).

³⁵ Gemma Fabregat Monfort, *Nuevas Perspectivas del Poder de Dirección y Control del Empleador* (Editorial Bomarzo 2016).

and public domains. Court analytics and healthcare analytics are very different from judge analytics and physician analytics respectively. If we permit the latter, there might be unpredictable outcomes seriously clashing with—and producing clashes among—human rights of different groups of stakeholders.

Within the justice context, these problematic issues stem from the justice system being one of the pillars of our democracies and one of the foundational stones of rule of law-based societies, besides being indispensable for citizens' trust in public institutions. Justice is also the only exclusively public domain that I address,³⁶ and this is the domain where the voices of disagreement arose most loudly, particularly in France.

However, the second domain that I address—namely, healthcare—is not less important and has its own particularities that make it special in terms of AI-based data analytics. The specificity lies in the powers that the patients have, that is a right to be part of decision-making processes, and in overall interaction with the healthcare system, which not even individuals in judicial proceedings have.

In addition, in both domains, data analytics put the professional between the hammer and the anvil, where the hammer is represented by the employer and the anvil by citizens. From this perspective, the question is whether the availability of data and possibility to measure and calculate not only performance but also to extract further new data out of it is desirable at any cost. The only shield the professionals have is provided by the rights guaranteed by data protection laws, yet the question is whether these laws are protective enough. The French legislator used them to ensure that French judges would not be subject to analytics. However, the question remains open and strikingly, other EU countries and the rest of the world appear not to follow the French example.

As to future research, there is an ethical—and to some extent also legal—question concerning private sector involvement in physician analytics. In particular, when it comes to healthcare, physician analytics could suggest or induce the patient to think (and consequently, to act) in a way that she would not have acted had she not had access to these analytics. This matters especially when such decisions are not beneficial for the patient and are not necessarily truthful. For instance, discovering a higher rate of complications in female patients of a particular surgeon might induce the patient to reject the option of surgery, when the surgery might be necessary or at least highly recommended. The complications at stake might not necessarily be caused by the physician, but depend on external factors, for instance, health conditions of patients—some surgeries are performed on very weak patients—or other circumstances. Perhaps this is a question for business and research ethics, but it is too important to be left for businesses themselves to deal with at their own discretion. We have already seen that many high-tech companies

³⁶ Private institutions, such as private arbitration, can also be subject to analytics, but I do not explore this topic here. On international arbitration in this sense, see Langford and others (n 7).

ignore the negative effects their products have on their customers, even on the particularly vulnerable ones: what matters are sales and *likes*, not human lives, as we have seen in the case of Facebook or Instagram and their effects on teenage girls.³⁷ Why should data analytics products be different and what can we, as a society, as academics, and as citizens, do to make them different?

Another interesting question for future research is how the prediction of judge (or other professional) performance by the analytics system impacts on her behaviour: if the system predicts judge's decision and the judge learns about it before the start of the trial, could it affect the decision of the judge and negatively impact her impartiality? Bufithis sees it as 'an intent to "psychologically force" judges to always go in the same direction'.³⁸ Although this question is not entirely legal but also belongs to psychology, it is still important to address.

That is all to say, data analytics on the one hand, and human rights—such as rights to privacy and personal data protection, non-discrimination, and equality, just to cite a few—on the other still have to find the equilibrium that would permit us to take the best of both: extract information out of data and use it to improve social services and strengthen democratic institutions, and at the same time, do so without paying a price in human rights.

³⁷ Georgia Wells, Jeff Horwitz, and Deepa Seetharaman, 'Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show' (*Wall Street Journal*, 14 September 2021) <<https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739>>.

³⁸ Gregory Bufithis, 'Understanding the French Ban on Judicial Analytics' (2019) <www.gregorybufithis.com/2019/06/09/understanding-the-french-ban-on-judicial-analytics/>.

Artificial Intelligence and the Right to an Effective Remedy

Sarah de Heer

1 Introduction

While the academic debate revolves mainly around the effects of artificial intelligence (AI) on fundamental rights,¹ specifically focusing on the right to non-discrimination,² its impact on the right to an effective remedy remains largely uncharted territory. This lacuna is not to be underestimated. While this right to an effective remedy is an important substantive right in itself, this right forms the cornerstone in protecting other substantive fundamental rights, which also includes the right to non-discrimination. Without an effective remedy, any recourse designed to uphold substantive human rights would be rendered futile. Thus, this contribution aims to fill this gap by examining the possible consequences of AI on the right to an effective remedy.

This chapter will first outline the international legal framework containing the right to an effective remedy (section 2). Subsequently, the chapter discusses the impact of AI applications on this right (section 3). Then the analysis shifts to the positive and negative implications of a specific AI application, namely automated decision-making (ADM) systems (section 4). Next, this chapter debates the regulatory solutions and possible policy reforms to mitigate the negative effects of ADM systems related to the opacity as regards how they work, and the processing of vast amounts of data (section 5).

¹ See eg Filippo A Raso and others, ‘Artificial Intelligence & Human Rights: Opportunities and Risks’ (2018) Berkman Klein Center for Internet and Society at Harvard University Research Publication <https://dash.harvard.edu/bitstream/handle/1/38021439/2018-09_AIHumanRights.pdf?sequence=1&isAllowed=y>; Max Vetz and Janneke Gerards, ‘Algoritme-Gedreven Technologieën en Grondrechten’ (2019) 21 Computerrecht 10.

² See eg Frederik Zuiderveen-Borgesius, ‘Strengthening Legal Protection Against Discrimination by Algorithms and Artificial Intelligence’ (2020) 24 International Journal of Human Rights 1572.

2 Legal Framework: The Right to an Effective Remedy

The right to an effective remedy is outlined in article 8 of the Universal Declaration of Human Rights (UDHR).³ This right has also been captured in binding international documents, including article 2(3) of the International Covenant on Civil and Political Rights (ICCPR),⁴ and in various regional human rights law instruments throughout the world,⁵ such as article 25 of the American Convention on Human Rights (ACHR),⁶ article 47(1) of the Charter of Fundamental Rights of the European Union (EU Charter),⁷ and article 13 of the European Convention on Human Rights (ECHR).⁸ The European Court of Human Rights (ECtHR) held that an effective remedy consists of ‘a remedy that is as effective as can be having regard to the restricted scope for recourse inherent in any system’.⁹ The ECtHR, thus, requires a remedy not only to be ‘effective’ in law but also in practice.¹⁰ Further, a remedy is only effective, provided it is adequate and accessible.¹¹

2.1 General Remarks

The right to an effective remedy and the right to a fair trial are closely connected.¹² The close ties are illustrated by the fact that the right to a fair trial emanates from the right to an effective remedy, as without a fair trial an effective remedy cannot be guaranteed.¹³ The difference between these two rights is the following: the right

³ Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR).

⁴ International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR).

⁵ Note that international environmental law also enshrines the right to an effective remedy. For example, art 9 of the Aarhus Convention, Principle 10 of the Rio Declaration, and art 235 of the United Nations Convention on the Law of the Sea. Further, while the African Charter on Human and Peoples’ Rights (ACHPR) does not explicitly mention the right to an effective remedy, the African Commission has given recognition to this right through its monitoring procedures, see Godfrey Musila, ‘The Right to an Effective Remedy Under the African Charter on Human and Peoples’ Rights’ (2006) 6 African Human Rights Law Journal 442. Likewise the Asian Human Rights Charter does not include the right to an effective remedy. However, art 5 of the Association of Southeast Asian Nations Human Rights Declaration, which is a supplement to the Charter, encompasses this right. It is noteworthy that the Asian Human Rights Charter is a people’s charter, which means that due to a lack of a government-issued human rights document, this charter was constructed in hopes to create a human rights friendly environment.

⁶ American Convention on Human Rights (22 November 1969), B-32 (ACHR).

⁷ Charter of Fundamental Rights of the European Union [2012] OJ C326/391 (EU Charter).

⁸ Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR).

⁹ *Klass v Germany* App no 5029/71 (ECtHR, 6 September 1978), para 69.

¹⁰ *Ilhan v Turkey* App no 22277/93 (ECtHR, 27 June 2010), para 97.

¹¹ *McFarlane v Ireland* App no 31333/06 (ECtHR 10 September 2010), para 108.

¹² For a comprehensive analysis of the dynamics between AI and fair trial rights specifically, see the chapter by Helga Molbæk-Steenisig and Alexandre Quemy in this volume.

¹³ Tobias Lock and Denis Martin, ‘Article 47 CFR: Right to an Effective Remedy and to a Fair Trial’ in Manuel Kellerbauer, Marcus Klamert, and Jonathan Tomkin (eds), *The EU Treaties and the Charter of*

to an effective remedy is a substantive right that ensures an adequate means of redress, while the right to a fair trial is a procedural right that safeguards the effective access to a fair hearing,¹⁴ which ensures access to judicial proceedings. The overall right to a fair trial consists of various conditions, one of which is securing the right to a fair hearing. Since especially the right to a fair hearing may be curtailed due to the use of generic AI applications and more particularly ADM systems, section 2.2 will zoom in on its requirements.

2.2 The Right to a Fair Hearing

The right to a fair hearing is crucial to realise the effective use of the right to an effective remedy, since the right to a fair hearing largely depends on whether an individual can meaningfully rely on their right to an effective remedy. The right to a fair hearing is a threefold condition. First, a fair hearing requires ensuring the adversarial principle, or the principle of *audi alteram partem*, that calls for parties to the proceedings to be informed of, comprehend, and provide their comments on the case pending.¹⁵ This does not only cover the submissions of the opposing party, but also the pleas of law raised by the court on its own motion. The key element being that parties should be able to discuss all points introduced by both the opposing side and the court.¹⁶ In other words, the principle of *audi alteram partem* demands that parties effectively participate in the proceedings.¹⁷

Second, a fair hearing demands to safeguard the principle of equality of arms,¹⁸ or the principle of procedural equality, which is closely akin to the adversarial principle.¹⁹ While under the former principle parties are to be treated equally, the latter requires parties to the proceedings to have access to relevant materials irrespective of whether these materials are in fact available to the opposing party.²⁰ The principle of equality of arms embodies the idea of a fair balance between the parties.²¹ Thus, the parties in front of the court are to have the same procedural rights—unless

Fundamental Rights: A Commentary (OUP 2019) para 6; Nihal Jayawickrama, *The Judicial Application of Human Rights Law: National, Regional and International Jurisprudence* (CUP 2017) 493.

¹⁴ Dinah Shelton, 'Sources of Article 47 Rights' in Steve Peers and others (eds), *The EU Charter of Fundamental Rights: a Commentary* (Hart 2015) para 47.10.

¹⁵ Case C-89/08 *European Commission v Ireland* (ECJ, 2 December 2009), para 52. See also art 2(e) of the Principles and Guidelines on the Right to a Fair Trial and Legal Assistance in Africa, 2003, DOC/OS(XXX)247.

¹⁶ Case C-471/11 *Banif Plus Bank* (ECJ, 21 February 2013), paras 29–31.

¹⁷ Bernadette Rainey, Elizabeth Wicks, and Clare Ovey, *The European Convention on Human Rights* (OUP 2017) 296.

¹⁸ See also art 2(a) of the Principles and Guidelines on the Right to a Fair Trial and Legal Assistance in Africa, 2003, DOC/OS(XXX)247.

¹⁹ Lock and Martin (n 13) para 29.

²⁰ *Nideröst-Huber v Switzerland* App no 18990/91 (ECtHR, 18 February 1997), paras 23–24.

²¹ *Dombo Beheer BV v the Netherlands* App no 14448/88 (ECtHR, 27 October 1993), para 33.

there is an objective and reasonable justification—and neither party is to be placed in a significant disadvantage.²² Under the principle of equality of arms the judiciary needs to grant the parties to the proceedings a reasonable opportunity to provide their points of view on the submissions of the opposing party.²³ Moreover, the parties are to be given the occasion to be represented by a legal counsel.²⁴ Further, the judiciary may not use submissions that the opposing party has not been able to examine in its reasoning.²⁵ The principle of procedural equality includes a right to be heard—especially when adopting an unfavourable decision²⁶—which means that parties are to plead their case under circumstances that are not less favourable compared to the opposing party.²⁷ This principle is hampered when one party to the proceedings has experienced serious practical obstacles to plead its case.²⁸ Such a barrier to the principle of equality of arms appears to be the case when a party in front of the court is not given permission to hold an opening and a closing speech and is not allowed to make submissions on matters of law.²⁹ Generally, the principle of equality of arms has been violated in case the overall proceedings have been unfair.³⁰

Third, a fair hearing also includes the right to a reasoned judgment, which requires the judiciary to deliver a reasoned ruling that enables the parties to the proceedings to understand the court's reasoning and to decide whether they should appeal its judgment.³¹ The court should clearly mention the grounds upon which it has based its ruling.³² Even though a court is not obliged to discuss each argument submitted by the parties,³³ this is not to be construed as a *carte blanche* to

²² Sangeeta Shah, 'Detention and Trial' in Daniel Moeckli and others (eds), *International Human Rights Law* (OUP 2014) 274.

²³ Case C-199/11 *European Community v Otis NV* (ECJ, 6 November 2012), para 71.

²⁴ *Ntukidem v Oko* App no SC.30/1989 (Supreme Court of Nigeria, 12 February 1993). See also art 7(1)(c) of the ACHPR; art 47(2) of the EU Charter; and art 2(f) of the Principles and Guidelines on the Right to a Fair Trial and Legal Assistance in Africa 2003, DOC/OS(XXX)247.

²⁵ *Ernst v Belgium* App no 33400/96 (ECtHR, 15 July 2003), paras 60–61.

²⁶ *Holland v Minister of the Public Service, Labour and Social Welfare* App no ZLR 186 (S) (Supreme Court of Zimbabwe). The ECtHR also noted that public administration not giving the individual—a company—the opportunity to be heard is problematic in light of procedural safeguards, see *Megadat.com SRL v Moldova* App no 21151/04 (ECtHR, 8 April 2004), para 73.

²⁷ Case C-169/14 *Sánchez Morcillo* (ECJ, 17 July 2014), para 49.

²⁸ *Makhfi v France* App no 59335/00 (ECtHR, 19 October 2004), para 40.

²⁹ *Hurnam v Paratian*, Privy Council on appeal from the Court of Civil Appeal of Mauritius [1998] 3 LRC 36.

³⁰ Debbie Sayers, 'Article 47(2): Everyone is Entitled to a Fair and Public Hearing Within a Reasonable Time by an Independent and Impartial Tribunal Previously Established by Law. Everyone Shall Have the Possibility of Being Advised, Defended and Represented' in Steve Peers and others (eds), *The EU Charter of Fundamental Rights: a Commentary* (Hart 2015) para 47.208. For example, a violation of the principle of equality of arms may be remedied by an appeal court, and thus the overall proceedings may be deemed fair. See *Schuler-Zgraggen v Switzerland* App no 14518/89 (ECtHR, 24 June 1993), para 52.

³¹ Case C-283/05 *ASML Netherlands BV* (ECJ, 14 December 2006), para 28. See also art 2(i) of the Principles and Guidelines on the Right to a Fair Trial and Legal Assistance in Africa 2003, DOC/OS(XXX)247.

³² *Hadjianastassiou v Greece* App no 12945/87 (ECtHR, 16 December 1992), para 33.

³³ Rainey, Wicks, and Ovey (n 18) 294.

disregard persuasive and relevant arguments that would affect the outcome of the proceedings.³⁴ These arguments need to be addressed specifically and expressly by the court.³⁵ The extent of the judiciary's duty to provide reasons for its judgments hinges on the nature of the decision, which the court should examine by considering the proceedings as a whole and the circumstances of the decision. Concisely, the court should consider whether the procedural safeguards surrounding the decision ensure that the parties to the proceedings are able to lodge an appeal.³⁶

3 AI Applications and the Right to an Effective Remedy

3.1 Generic AI Applications

The consequences of AI applications on the right to an effective remedy may be specifically alarming. Indeed, individuals maintaining that their substantive human right—ranging from the right to freedom of expression and the right to privacy—have been hampered, will rely on their right to an effective remedy to undo the alleged violation. As a result, hindering the right to an effective remedy may lead to fundamental rights no longer being effectively safeguarded. Nevertheless, both private industry and public administration use AI applications that potentially affect the right to an effective remedy—and consequently the substantive human right.

Turning to the private sector that has long ago fully embraced AI applications in their daily business, the following example shows the issues around the right to privacy. Watrix, a Chinese company, has successfully developed gait recognition technology, which can recognise individuals by analysing their body shape and how their arms move with a 96 per cent accuracy.³⁷ Another illustration may serve to show obstruction to the right to property. Zillow, an American online real estate enterprise, uses an algorithm in its Zestimate program that provides price estimates of properties by looking at, amongst others, photos of the property and how many days the property has been listed.³⁸

Now moving to the public sector that has not shunned from the use of AI applications. In these cases, the greatest concern seems to revolve around the right to privacy and the right to data protection. In Kenya, citizens and residents are to provide the government their biometric data, including earlobe geometry and voice

³⁴ *Dhahbi v Italy* App no 17120/09 (ECtHR, 8 April 2014), paras 31–33.

³⁵ *Ruiz Torija v Spain* App no 18390/91 (ECtHR, 9 December 1994), para 30.

³⁶ Case C-619/10 *Trade Agency Ltd* (6 September 2012), para 60.

³⁷ ‘There’s Facial Recognition, and Now, Watrix Technology’s Gait Recognition System That Recognizes Your Walk’ (*Techeblog*, 27 February 2019) <www.techeblog.com/watrix-technology-gait-recognition-system/>.

³⁸ Stan Humphries, ‘Introducing a New and Improved Zestimate algorithm’ (*Zillow*, 27 June 2019) <www.zillow.com/tech/introducing-a-new-and-improved-zestimate-algorithm/>.

recordings, to obtain a national ID.³⁹ Another illustration stems from Chile, where the National Board of Scholarships and School Aid uses a facial recognition tool to distribute school meals.⁴⁰

These generic AI applications used by both private sector and public administration also have an impact on an individual's right to an effective remedy. The question rises how an individual ought to seek redress for impaired substantive rights with limited or no knowledge on how these generic AI applications operate.

3.2 Specific AI Application: ADM Systems

One specific AI application may particularly affect the right to an effective remedy of an individual, namely ADM systems, which are algorithmic tools that predict an outcome. The outcome of ADM systems is either a full decision or a partial decision, which is aimed at facilitating the tasks of the users, in this case, companies or public administration. While full automation systems issue full decisions that thus lack further human involvement, the use of partial automation systems results in partial decisions that still require further human involvement. In other words, full automation takes over the whole decision-making process as opposed to partial automation that merely replaces certain parts thereof.⁴¹

First, the private industry, where routine ADM systems may have a disturbing effect on substantive human rights. Instagram's use of algorithms may hamper the right to freedom of expression, more specifically the right to impart information. Instagram uses AI to proactively identify and remove content that may be in violation of Instagram's Community Guidelines—and thus before an individual has reported the allegedly incompatible content.⁴² Another example is painted against the backdrop of the right to non-discrimination. Since 2014, Amazon had been creating an algorithm to assist with its recruitment procedure. However, Amazon decided to halt this project in 2018, since its algorithm showed a negative

³⁹ Amnesty International, 'Amnesty International Submission to the Office of the United Nations High Commissioner for Human Rights on the Impact of Digital Technologies on Social Protection and Human Rights' (Office of the United Nations High Commissioner for Human Rights) <www.ohchr.org/sites/default/files/Documents/Issues/Poverty/DigitalTechnology/AmnestyInternational.pdf>.

⁴⁰ Diego Bastarrica, 'Chile: Children Asked for Their Biometrics Data to Obtain Food Rations in Schools' (*Privacy International*, 17 February 2019) <<https://privacyinternational.org/examples/2880/chile-children-asked-their-biometrics-data-obtain-food-rations-schools>>.

⁴¹ Michael Vaele and Irina Brass, 'Administration by Algorithm? Public Management Meets Public Sector Machine Learning' in Karen Yeung and Martin Lodge (eds), *Algorithmic Regulation* (OUP 2019) 123–27.

⁴² 'How Does Instagram Use Artificial Intelligence to Moderate Content?' (*Help Centre*) <https://help.instagram.com/help/instagram/423837189385631/?locale=en_GB&maybe_redirect_pol=true>.

bias towards female applicants, which meant that the algorithm favoured male candidates.⁴³

Second, focusing on public administration, where there are a myriad of examples demonstrating that governments are rapidly moving towards the adoption of partial—or even full—automated decisions. A first example is located in Canada and may have detrimental outcomes for the right to asylum. Based on the Immigration and Refugee Protection Regulations, algorithms are used to establish the so-called ‘safe countries’ on the ‘Designated Countries of Origin’ list. The algorithm determines whether a country is ‘safe’ by considering, amongst others, the number of refugee status that has been granted by the government.⁴⁴ Another illustration concerns the right to social security and is situated in Australia, where the implemented Targeted Compliance Framework requires job seekers to report mutual obligation requirements and to check their compliance status on an online dashboard. Failing to meet a mutual obligation may lead to a demerit—without any human involvement. Each demerit expires after six months. More than five demerits may eventually result in postponed payments or even financial penalties after human involvement.⁴⁵

As demonstrated by the two examples above, governments have created their own ADM systems. Nonetheless, they largely lack the required technical expertise to build, maintain and operate these systems and they thus venture to the private industry that has ample technical knowledge.⁴⁶ The reliance on the private sector can be divided in three categories, namely: (i) acquiring commercially built ADM systems; (ii) procuring operational and/or maintenance ADM systems-related services; and (iii) data sharing between public administration and private parties. The following two examples serve to show impediments to the right to social security. The first illustration concerns Canada, where public administration has bought a commercial ADM system. The financial aid programme of the Province of Ontario, ‘Ontario Works’, uses Social Assistance Management System (SAMS) to decide which individuals should be granted benefits. SAMS is based upon Cúram, a customisable software package from third party, IBM.⁴⁷ The second example is

⁴³ Jeffrey Dastin, ‘Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women’ (*Reuters*, 11 October 2018) <www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.

⁴⁴ Petra Molnar and Lex Gill, *Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada’s Immigration and Refugee System* (University of Toronto Press 2018) 33.

⁴⁵ Secretary General, ‘Report of the Special Rapporteur on Extreme Poverty and Human Rights’ (11 October 2019) UN Doc A/74/48037, para 31; Australian Government, ‘Targeted Compliance Framework’ (Department of Education, Skills and Employment) <<https://www.workforceaustralia.gov.au/content/documents/individuals/obligations/learn/compliance-demerits/targeted-compliance-framework-brochure.pdf>>.

⁴⁶ Molnar and Gill (n 45) 60.

⁴⁷ Human Rights Watch, ‘May 2019 Submission to the UN Special Rapporteur on Extreme Poverty and Human Rights Regarding His Thematic Report on Digital Technology, Social Protection and Human Rights’ <www.ohchr.org/sites/default/files/Documents/Issues/Poverty/DigitalTechnology/HumanRightsWatch.pdf>; Secretary General (n 46) para 21.

situated in the Netherlands, where government uses the services of an enterprise. The municipality of Nissewaard relied on the services provided by Totta data lab to predict which citizens may commit social assistance fraud.⁴⁸ Because of the uproar in the media and amongst the public, Nissewaard ceased the use of Totta data lab's services at the beginning of July 2021.⁴⁹

Concerns related to another fundamental right may be witnessed in, for example, Poland, where data sharing amongst the private industry and the public administration may thwart the right to data protection. The Polish Tax Authority uses an ADM system, Clearance Chamber ICT System, that collects data from various financial institutes, including banks and credit unions. Based on the data gathered, this ADM system suggests to the Tax Authority which accounts are susceptible of fraudulent activities.⁵⁰

The above examples illustrate how various human rights may be curbed due to ADM systems used by either companies or the public sector. The individual may not understand how the ADM system has reached its decision that has negatively affected their human rights. As a result, the individual is not in a position to effectively oppose the automated decision, irrespective whether it concerns a partial or full decision. Further, the judiciary is faced with the strenuous task of safeguarding the right to an effective remedy, as the court may be equally left in the dark as regards the modus operandi of the ADM system that has issued the full or partial automated decision. The following section addresses the negative consequences experienced by individuals when relying on their right to an effective remedy.

4 The Positive and Negative Consequences of ADM Systems on the Right to an Effective Remedy

4.1 General Characteristics of ADM Systems

The use of ADM systems brings about their own unique threats, which includes the obscurity surrounding fathoming their modus operandi and their ability to process vast amounts of data. The challenges by virtue of obscurity stem from different causes. First, data engineers may intentionally build the underlying model of ADM systems to create opacity. This is the so-called 'black box' and may originate from employing machine learning techniques while designing the underlying code of

⁴⁸ Judith Nuijens, 'Bijstandsfraude Voorspellen Met Big Data' (*Sociaal Web*, 2017) <<http://magazines.sociaalweb.nl/fraude#!/bijstandsfraude-voorspellen-met-big-data>>.

⁴⁹ Redactie Economie, 'Nissewaard Stopt Met Omstreden Methode om Fraude Met Bijstand op te Sporen' (*Trouw*, 7 July 2021) <www.trouw.nl/economie/nissewaard-stopt-met-omstreden-methode-om-fraude-met-bijstand-op-te-sporen~b985a30d/>.

⁵⁰ Natalia Mileszyk and Alek Tarkowski, 'Automating Society Report: Research—Poland' (*Algorithm Watch*, 2020) <<https://automatingsociety.algorithmwatch.org/report2020/poland/>>.

the model of ADM systems. As a result of these techniques, the model may show autonomous behaviour. Put differently, such ADM systems possess the ability to adapt and learn independently, namely without human involvement.⁵¹ Hence, data scientists—let alone the public—are faced with an impossible task, namely comprehending how the algorithm reached the outcome based on the input data. Second, even if the algorithms are not intentionally created to be ungraspable, opaqueness may stem from simply the complexity of the code underlying the model of ADM systems.⁵² In other words, designing and understanding algorithms may require technical expertise. Since the general public generally lacks this level of knowledge due to technical illiteracy, unravelling how the model reaches its outcome may—practically—be an impossible task for the average person. Third, the models used for these ADM systems may be the key to success of the enterprise. The underlying model may be the company's competitive advantage, which makes their business profitable. Hence, these private parties may want to rely on intellectual property rights to prevent their competitors from copying their code and thus benefiting as a result from their technical expertise. Therefore, enterprises in the possession of such a lucrative model may choose to use the safeguards offered by, for example, the regime of trade secrets,⁵³ which bars making the code known. As a result, this practice may exacerbate the opaqueness caused by the 'black box' and the complexity of ADM systems, since this protection mechanism creates an additional hurdle to fathom how the model has reached the outcome.

The automatic nature of ADM systems allows the processing of immense volumes of data. Seeing the processing capacities of such ADM systems, challenges of inaccurate and incomplete data are more likely to occur. Two trends have amplified these complexities arising due to the ADM systems' ability to process large quantities of data, namely ADM systems may also facilitate (i) combining datasets amongst companies and public administration; and (ii) data sharing amongst public authorities and the private industry.

4.2 Positive and Negative Consequences

While ADM systems have positive effects on the right to an effective remedy, it seems challenging to reconcile them to the three requirements of the right to a fair hearing, namely (i) the adversarial principle; (ii) the principle of equality of arms; and (iii) the right to a reasoned judgment. Nevertheless, various ADM systems may have a positive outcome on the other elements of the right to a fair trial.

⁵¹ Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society 1, 4–5.

⁵² *ibid* 4.

⁵³ *ibid* 3–4.

One example includes law firms using ADM systems to predict the likelihood their case may have a positive outcome.⁵⁴ Since lawyers and their clients may choose not to litigate cases with an off-chance of winning, the courts end up having more time to spend on the more complicated cases, which benefits the criterion of 'reasonable time'. Another illustration revolves around DoNotPay, a so-called 'robot lawyer'. DoNotPay acts as a lawyer by supporting individuals who bring a case to court. Their services are rendered for only a portion of the average, regular legal fees. DoNotPay may improve the requirement of 'the right to access to a court', since the regular costly hourly rates may form an obstacle.

The negative impact of ADM systems on the right to an effective remedy, and more specifically the threefold criteria of the right to a fair hearing are caused due to the opaqueness of ADM systems—potentially worsened by the protection mechanisms provided by intellectual property rights, and their ability to process a large amount of data—possibly exacerbated by simplifying the practice of merging datasets and data sharing.⁵⁵ The obscurity stemming from the 'self-learning'-nature and the complexity of ADM systems may lead to hampering the right to a fair hearing. More specifically, the question rises how the judiciary can safeguard the adversarial principle that requires parties to effectively partake in the proceedings, when not all parties to the proceedings comprehend how the algorithm underlying the ADM system works. Further, courts may struggle to ensure the principle of equality of arms, since parties may not be given a reasonable opportunity to provide their points of view on the opposing party's submissions due to their lack of understanding of how the ADM system reached its decision. Moreover, since judges may struggle to grasp the operation of the ADM system, they may not be able to hand down a reasoned ruling that allows the parties to decide whether to appeal the ruling. Concisely, the three elements of the right to a fair hearing may be at risk, since fathoming the modus operandi of the algorithm underlying the ADM systems may be highly cumbersome or impossible for not only the parties to the proceedings but also the judiciary. The protection mechanisms under intellectual property rights forms an additional hurdle created by the opacity of ADM systems. While how they work may already be—nearly—impossible to unravel, intellectual property rights bar that the model is made public, which further complicates the model's understanding.

⁵⁴ Alex Heshmaty, 'Use of AI in Law Firms to Predict Litigation Outcomes' (*LexisNexis*, 1 February 2022) <www.lexisnexis.co.uk/blog/future-of-law/using-ai-to-predict-litigation-outcomes>.

⁵⁵ Further, decisions issued by ADM systems may be flawed owing to the prejudices built into the ADM systems and due to the input of biased data. As a result, the right to non-discrimination may be especially at risk, since the automated decision may have an adverse impact on certain segments of society. However, an outline of the consequences of the prejudices inherent to ADM systems and the input of biased data is not the main aim of this contribution's discussion. For an overview of prejudices inherent to ADM systems, see Shahriar Akter and others, 'Algorithmic Bias in Data-Driven Innovation in the Age of AI' (2021) 60 *International Journal of Information Management* 102387.

The possibility of ADM systems to process vast amounts of data of individuals may also hamper the three elements of the right to a fair hearing. While parties to the proceedings could be provided with the complete dataset that has led to the impugned decision, the question rises whether they are capable of comprehending these large datasets. As a result, these complexities surrounding the extensive nature of the datasets may pose a risk to the adversarial principle, since the parties in front of the court may be prevented from effectively taking part in the judicial proceedings. The principle of equality of arms may also be thwarted, as the parties may not be able to put forward their points of view due to their inability to grasp immense datasets. Further, the judiciary may be faced with the cumbersome task of unravelling such extensive datasets, which may result in hampering the right to a reasoned judgment that enables the parties to the proceedings to decide whether or not they should appeal the decision. Moreover, employing large quantities of data may lead to the possible inclusion of inaccurate and incomplete data, which may have unfavourable effects on individuals. These two characteristics are only amplified by the relatively new trends of data merging and data sharing, as the potential datasets will become even more extensive.

Consequently, due to these obstacles posed by ADM systems used by both the private and public sector, an effective remedy is no longer accessible to individuals whose fundamental rights are adversely affected by such systems. The remedy is thus no longer effective in practice. Put differently, the right to an effective remedy is rendered meaningless in the context of ADM systems.

5 Regulatory Solutions and Possible Solutions

5.1 ADM Systems: Obscurity

To address the obscurity due to complexity, a distinction needs to be drawn between the use of ADM systems by on the one hand the public sector and on the other hand the private sector. Public administration may opt to solely use open-source ADM systems, whose publication is thus not barred due to intellectual property rights. The open-source model could be added in an algorithm register that lists all systems employed by public administration. These algorithm registers are already in place in the municipalities of Amsterdam⁵⁶ and Helsinki.⁵⁷ These registers provide the following information on the ADM systems used by these two municipalities, namely: (i) the datasets employed during the development phase and the utilisation phase; (ii) explanations as to how the model works; (iii) how

⁵⁶ ‘City of Amsterdam Algorithm Register Beta’ (*Gemeente Amsterdam*) <<https://algoritmeregister.amsterdam.nl/en/ai-register/>>.

⁵⁷ ‘City of Helsinki AI Register’ (*Helsinki*) <<https://ai.hel.fi/en/ai-register/>>.

the right to non-discrimination is safeguarded; (iv) whether the system is placed under human oversight; and (v) the risks associated with the use of the system and how they are managed. This enables the general public to understand how the model works, since the explanations clarify how the model reached the outcome. Turning towards the private industry, such an algorithm register would only entail a solution for the companies that are willing to make their ADM systems open-source, which may be impossible due to intellectual property rights. In these cases, companies ought to publish additional explanations to understand the operation of ADM systems. Since the parties to the proceedings and the judiciary would be able to fathom the ADM system's reasoning thanks to the algorithm register or further clarifications, this would allow proceedings to be compliant with the adversarial principle, the principle of equality of arms, and the right to a reasoned judgment. Nonetheless, the algorithm register and further clarifications would not aid the comprehension of both public and private ADM systems that are opaque due to the black box, as explaining the modus operandi of such models remains inconceivable—even for data scientists. Thus, the question arises if such models should be used.

In vertical relationships, governments are under an obligation to safeguard human rights, including the right to an effective remedy. This requirement is not imposed on private parties in horizontal relationships.⁵⁸ However, a change is ongoing: the United Nations Guiding Principles on Business and Human Rights (UNGPs) point at the need for the private industry to respect human rights. Specifically, UNGP 13 requires companies to avoid or prevent causing harm to human rights. To this end, the private sector needs to identify these human rights risks. To do so, UNGPs 17–19 call for businesses to conduct human rights due diligence, which should include a human rights impact assessment. Based on these UNGPs, the private sector ought to assess the impact of their ADM systems on the adversarial principle, the principle of equality of arms, and the right to a reasoned judgment. Even though performing a human rights impact assessment is currently only imposed within horizontal relationships, nothing bars the legislator to also include such an obligation for public authorities that use ADM systems in their daily activities.⁵⁹

⁵⁸ Human rights under the European Convention on Human Rights may have indirect horizontal effect by imposing positive obligations upon States to protect human rights in horizontal relationships and by requiring national courts to apply human rights in horizontal disputes, see Janneke Gerards, *General Principles of the European Convention on Human Rights* (CUP 2019) 144–59.

⁵⁹ The requirement to conduct such an impact assessment upon both public and private parties is no novelty. Indeed, the GDPR also requires the performance of an impact assessment to determine which processing activities may pose a high risk to the rights of data subjects, see art 35.

5.2 ADM Systems: The Ability to Process Immense Amounts of Data

Zooming in on the EU, the General Data Protection Regulation⁶⁰ (GDPR) may curb the challenges resulting from the vast amount of data processed on individuals that are potentially worsened due to the practice of data merging and data sharing. Under the GDPR, the processing of personal data⁶¹ needs to adhere to various principles that aim at quantitatively restricting the amount of data being processed. For example, the principle of purpose limitation⁶² only permits the processing of personal data that is necessary for the purpose for which the ADM systems are used. The principle of data minimisation⁶³ requires that data processing is limited to personal data that is necessary to fulfil the purpose pursued by the ADM system. Under the principle of data accuracy only accurate and up-to-date personal data may be processed in light of the purpose of the ADM system.⁶⁴ Put differently, the principle of purpose limitation, the principle of data minimisation and the principle of accuracy may help to overcome the issues of unlawfully processing personal data, processing more personal data than strictly required for the purpose at stake, and the processing of inaccurate data. Taking into account these principles will inevitably restrict the amount of data processed by ADM systems, and may thus lead to a smaller dataset. As a result, both the parties to the court's proceedings and the judiciary will be faced with a more feasible task. Put differently, the three elements of a fair hearing, namely (i) the adversarial principle; (ii) the principle of equality of arms; and (iii) a reasoned ruling, are more likely to be safeguarded when the principle of purpose limitation, the principle of data minimisation and the principle of data accuracy are respected. Since the GDPR has already been effective for about five years, the above-mentioned solutions are nothing new. However, since legal compliance and enforcement of the GDPR are lagging behind, this chapter calls for a more prudent adherence to the safeguards in the GDPR.

6 Conclusion

This contribution explored the use of AI applications—more specifically ADM systems—against the backdrop of the right to an effective remedy, which forms

⁶⁰ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (GDPR).

⁶¹ The notion of 'personal data' under art 4(1) of the GDPR is broad. See eg Case C-582/14 *Patrick Beyer v Bundesrepublik Deutschland* (ECJ, 19 October 2016); Case C-434/16 *Peter Nowak v Data Protection Commissioner* (ECJ, 20 December 2017).

⁶² GDPR, art 5(1)(b).

⁶³ ibid art 5(1)(c).

⁶⁴ ibid art 5(1)(d).

the cornerstone in guarding substantive human rights. However, the individual's right to adequately make use of an effective remedy is contingent on the right to access a court, which the right to a fair trial—and particularly the right to a fair hearing—aims to facilitate. Thus, this piece examined the three requirements of the right to a fair hearing, which are (i) the adversarial principle; (ii) the principle of equality of arms; and (iii) a reasoned ruling. The hurdles posed by ADM systems upon these three criteria cannot be underestimated. The use of these ADM systems curtails the right to a fair hearing due to the specific characteristics of ADM systems, which include their obscurity—potentially worsened by intellectual property rights—and their ability to process large quantities of data—potentially exacerbated by the practice of data merging and data sharing—which pose a threat to especially the three requirements of the right to a fair hearing. Without a fair hearing, an individual can hardly be deemed to meaningfully exercise their right to an effective remedy. As a result, ADM systems may seriously hamper the right to an effective remedy, which means that there is no legal redress for alleged violations of substantive fundamental rights. While there are regulatory and policy frameworks in place to overcome the challenges posed by the typical features of ADM systems, including an algorithm register and the GDPR, it is no clear-cut task to overcome these characteristics. Thus, it is now time to examine how to ensure the right to an effective remedy in light of these ADM systems.

PART VI

ARTIFICIAL INTELLIGENCE
AND ASYLUM

21

Artificial Intelligence Technologies and the Right to Seek and Enjoy Asylum

An Overview

Raimy Reyes

1 Introduction

When individuals do not have their rights guaranteed by their own state, they have the internationally recognised right to flee their countries and ‘to seek and to enjoy in other countries asylum from persecution’.¹ This process of forced migration can be complex and cumbersome—it starts with the decision to leave one’s home country, travelling by different means to another country, accessing the territory, filing for asylum there, and ultimately being recognised as a refugee. In 2022, the United Nations (UN) Refugee Agency (United Nations High Commissioner for Refugees or UNHCR) estimated the total number of people worldwide who were forced to flee their homes due to conflicts, violence, and human rights violations to exceed 100 million people, reaching record high levels.²

The reality of forced migration and the guarantee of the right to asylum is directly linked with international obligations by host countries to adapt their legal framework and administrative procedures to receive people forced to flee,³ process their asylum claim, and—when appropriate—recognise them as refugees. States, international organisations, and other stakeholders in migration management have already started to explore the use of artificial intelligence (AI) technologies to manage people forced to flee—despite the lack of a proper international and/or domestic legal framework to guarantee the right to seek asylum and, moreover, ignoring the real impact on human lives the improper use of AI can have.

AI technologies have an appeal as tools to be used to manage forced migration. These new technologies are already implemented by many states through

¹ Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR), art 14.

² UNHCR, ‘Global Trends: Forced Displacement in 2021’ (2022) 7.

³ The term ‘people forced to flee’ will be used throughout this chapter to encompass asylum seekers, refugees, and others who have not filed for asylum or been recognised as refugees but are in need of international protection.

automated facial recognition (AFR) systems, ground sensors, aerial video surveillance drones, biometric databases, automated decision-making (ADM), and many other aspects of migration management. Ultimately, the primary purpose of the technologies used in migration management is to track, identify, and control those crossing borders. Since these technologies were not originally designed to benefit those forced to flee, their main focus is not to uphold their human rights.

The use of AI could certainly have a positive impact in some limited scenarios when properly regulated and implemented. However, states cannot relegate their international human rights obligations solely to AI technologies. Inherently biased data and the use of AI can have a detrimental impact and lead to the breach of multiple international obligations when applied to life-or-death decision-making procedures such as access to territory, filing for asylum, the right to have a fair and impartial decision-maker on asylum claims, and being able to appeal negative decisions. Unfortunately, the race to develop AI is unlikely to be stopped or slowed down by international human rights obligations, particularly since these technologies are developed by the private sector and implemented on an already marginalised population, without oversight or accountability.

2 AI Technologies and Forced Migration: Positive and Negative Implications

The right to seek asylum has a broad reach and is fundamentally linked to due process guarantees at multiple stages traditionally upheld by administrative authorities such as immigration agents, asylum adjudicators, and administrative judges. To understand where AI is relevant to the right to seek asylum, it is best to look through the lens of the forced migration journey: the decision to flee and the right to access the territory of asylum (section 2.1); the right to file for asylum (section 2.2); the refugee status determination process (section 2.3); and the durable solutions that end the displacement cycle (section 2.4).

Through these stages of the forced migration journey, human rights bodies have established some of the minimum procedural guarantees for people seeking international protection:⁴

- (i) access to territory and to not be forcedly returned;
- (ii) authorities trained to identify international protection needs at borders;
- (iii) the right to be heard through a personal interview;

⁴ Inter-American Court of Human Rights (IACtHR), 'Journal of Jurisprudence No 2 People in Migration or Refugee Situation' (2020); Council of Europe, 'Guide on the Case-Law of the European Convention on Human Rights: Immigration' (ECHR, April 2022); Inter-American Commission on Human Rights (IACHR), 'Due Process in Procedures for the Determination of Refugee Status and Statelessness and the Granting of Complementary Protection' (OEA/Ser.L/V/II. Doc.255/20, 2020).

- (iv) the right to an interpreter;
- (v) confidentiality and the protection of personal data and information;
- (vi) Sharing the burden of proof and receiving the benefit of the doubt;
- (vii) the right to a reasoned, and substantiated decision;
- (viii) notification of decision;
- (ix) the right to a suitable and effective remedy;
- (x) remaining in the territory of asylum until a final decision has been reached;
- (xi) and a reasonable duration of the process.

In addition, states still have the duty to respect and guarantee all the other fundamental rights to people forced to flee.

As detailed below, some AI applications could assist in upholding some of the guarantees of the right to seek and enjoy asylum when properly regulated. However, it is important to keep in mind that the use of these emerging AI technologies can have far-reaching negative impacts in life-or-death decision-making procedures depending on how, and by whom, they are implemented. These technologies have been traditionally developed by the private sector for surveillance control purposes and have historical antecedents in colonial technologies of racialised governance. Thus, these AI technologies are not neutral and their design and use typically reinforce dominant social, political, and economic trends.⁵ Scholars have argued that the current lack of regulation of AI technologies in migration management is deliberate, as states single out non-citizens within their borders as a viable testing ground for new technologies.⁶

2.1 The Decision to Flee and the Right to Access the Territory of Asylum

Forecasting forced displacement movements is important, as accurate predictions can help save lives by allowing governments, international organisations, and other stakeholders to conduct a better informed allocation of humanitarian resources and to adapt their services to guarantee the rights of the incoming population.⁷

According to the UN Office for the Coordination of Humanitarian Affairs (OCHA), humanitarian decision-makers have called for the increased use of

⁵ UNGA, ‘Report of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance: Digital Borders’, UN GAOR Seventy-fifth Session UN Doc A/75/590 (2020), s 3.

⁶ Petra Molnar, ‘Technology on the Margins: AI and Global Migration Management From a Human Rights Perspective’ (2019) 305 Cambridge International Law Journal 306.

⁷ Diana Suleimenova, David Bell, and Derek Groen, ‘A Generalized Simulation Development Approach for Predicting Refugee Destinations’ (2017) 7 Scientific Reports 13377.

predictive analytics to inform anticipatory action.⁸ Predictive analysis can be used within the forced migration context given its ability to track, predict, and analyse population movements. Predictive analytics tools can analyse diverse data sources, such as satellite imagery, social media, mobile phone and IP data, and economic, political, geographic, and meteorological information. This data can be used to predict who will move, in what numbers, and their displacement pattern.⁹

An example of predictive analytics implemented to forced migration is ‘Project Jetson’, the UNHCR’s first AI-based predictive analytic tool aimed at providing predictions on the movement of displaced populations within and outside of Somalia, forecasting future arrivals and population movements in each region.¹⁰ AI technologies like ‘Project Jetson’, when regulated and used adequately, can have positive results and life-saving effects, for example, by allowing neighbouring countries and the international community to organise their resources to aid the incoming population. However, the opposite can be said when these technologies are used to keep people forced to flee at bay. Digital technologies are already being deployed to advance xenophobic and racially discriminatory ideologies, prevalent due to widespread perceptions of refugees and migrants as threats to national security.¹¹

Individuals in need of international protection may not be rejected at the border or points of entry into the territory in question without adequate analysis of their protection requests. States have an international obligation to allow entry to their territory to people forced to flee so they can materialise their right to seek asylum, otherwise, states may be liable to a breach of the right to seek asylum and the principle of *non-refoulement* (forced return). *Non-refoulement* is the cornerstone of international protection for refugees, asylum seekers, and others in need of international protection since it prevents the expulsion or return of such persons to the borders of territories where their life or freedom is in danger.¹² It is considered a customary rule of international law¹³ and as *jus cogens*.¹⁴

Currently, many states have already incorporated the use of AI to their so-called ‘smart border’ policies and actions to patrol their borders. The existent border-focused AI technologies are varied and derive from existing surveillance tools, they

⁸ UN Office for the Coordination of Humanitarian Affairs (OCHA), ‘Peer Review Framework for Predictive Analytics in Humanitarian Response’ (Centre for HumData 2020) <<https://centre.humdata.org/predictive-analytics/>>.

⁹ Niamh Kinchin, ‘Technology, Displaced? The Risks and Potential of Artificial Intelligence for Fair, Effective, and Efficient Refugee Status Determination’ (2021) 5 *Law in Context* <journals.latrobe.edu.au/index.php/law-in-context/article/view/157>.

¹⁰ UNHCR, ‘Project Jetson’ (2017) <jetson.unhcr.org/>.

¹¹ UNGA (n 5).

¹² Convention Relating to the Status of Refugees (adopted 28 July 1951, entered into force 22 April 1954) 189 UNTS 137 (Refugee Convention), art 33.

¹³ UNHCR, ‘Declaration of States Parties to the 1951 Convention and or Its 1967 Protocol relating to the Status of Refugees’ (16 January 2002) UN Doc HCR/MMSP/2001/09.

¹⁴ *The Institution of Asylum, and its Recognition as a Human Right under the Inter-American System of Protection*, Advisory Opinion OC-25/18 (IACtHR, 30 May 2018) para 98; *Soering v the United Kingdom* App no 14038/88 (ECtHR, 7 July 1989).

focus on detecting human arrivals in remote border areas, biometric analysis, AFR technology, mobile phone tracking, and other uses alike. Ultimately, computers—not humans—make preliminary determinations about possible threats and how authorities should respond.¹⁵

Emerging digital technologies are being developed and deployed in ways that are uniquely experimental, dangerous, and discriminatory in the border and immigration enforcement context.¹⁶ For example, in the United States (US), AI systems complement border officials with scanner software to monitor more territory in less time and at lower cost than might be otherwise possible. When the system detects movement by people or vehicles, it alerts Border Patrol (USBP) agents to follow up.¹⁷ FRONTEX, the European Border and Coast Guard Agency, has been testing various predictive analytics and unpiloted AI-powered military-grade drones in the Mediterranean for the surveillance and interdiction of migrant vessels seeking European shores to file asylum applications.¹⁸

However, these technologies far from upholding the rights of people forced to flee, are used as deterrent for people to reach countries of asylum, and have significantly increased migrant deaths as they push migration routes towards more dangerous terrains away from the surveillance technology.¹⁹ AI applications have already been used to support illegal interdiction measures, both on land and at sea, or the forced returns of refugees and migrants without consideration of individual circumstances and without the possibility to apply for asylum or appeal. Interventions like these breach *non-refoulement* obligations and are aided by surveillance technologies.²⁰

Another example of the issues arising from border-focused AI technologies is *iBorderCtrl*, a European Union (EU)-funded project piloted in Greece, Latvia, and Hungary from 2016 to 2019. The AI screened incoming travellers at border checkpoints and operated as a ‘video lie detector’. Travellers were asked a series of questions from a computer-animated border guard, via webcam, and their micro-gestures were analysed to see if they were lying. Lie detectors are not admissible

¹⁵ Hannah Tyler, ‘Border Metrics: The Increasing Use of Artificial Intelligence in Border Zones Prompts Privacy Questions’ (*Migration Policy Institute*, 2 February 2022) <www.migrationpolicy.org/article/artificial-intelligence-border-zones-privacy>.

¹⁶ UNGA (n 5).

¹⁷ Tyler (n 15).

¹⁸ Petra Molnar, ‘The EU’s AI Act and its Human Rights Impacts on People Crossing Borders’ (*DoT. Mig In Brief*, June 2022) <www.bosch-stiftung.de/sites/default/files/publications/pdf/2022-06/The%20EU%20AI%20Act%20and%20Its%20Human%20Rights%20Impacts.pdf>; Petra Molnar, ‘Robots and Refugees: the Human Rights Impacts of Artificial Intelligence and Automated Decision-Making in Migration’ in Marie McAuliffe (ed), *Research Handbook on International Migration and Digital Technology* (Edward Elgar 2021) 136.

¹⁹ Geoffrey Alan Boyce, Samuel N Chambers, and Sarah Launiu, ‘Democrats’ “Smart Border” Technology Is Not a “Humane” Alternative to Trump’s Wall’ (*The Hill*, 11 February 2019) <<https://thehill.com/opinion/immigration/429454-democrats-smart-border-technology-is-not-a-humane-alternative-to-trumps/>>; Tyler (n 15).

²⁰ UNGA (n 5).

as evidence in court precisely because they do not work, particularly taking into account differences in communication, stressed or nervous people, and the effects of trauma on memory.²¹ AFR has been proven to be highly discriminatory and biased, particularly against racialised groups, as well as inaccurate and culturally insensitive.²² Similarly, in 2017, an algorithm used by the US Immigration and Customs Enforcement Agency (ICE), originally used to determine whether incoming migrants should be detained or let out on bond after being arrested, was changed by ICE to recommend detention in every case in order to comply with the Trump administration's policies.²³

For the effectiveness of the right to seek and enjoy asylum, international human rights standards already dictate that border migration authorities should not refuse claims without adequately examining situations that may give rise to international protection needs. International human rights bodies have reiterated that refusing access to territory to people seeking international protection may breach different international obligations related to *non-refoulement*; prohibition of collective expulsions; non-punishment for irregular entry, and non-detention of migrants.²⁴ To uphold these obligations, border authorities must be properly trained in refugee law and have appropriate skills for identifying international protection needs.²⁵ At present, states still fail to meet their international obligation to adequately train their border officials to properly screen individuals forced to flee at arrival, resulting in many instances of *refoulement*. It would pose a challenge for states to comply with the obligation to train border officials on asylum procedures if these duties are relegated to AI.

While use of AI could have positive impacts when used correctly to predict mass influxes of incoming population to tend to their needs and aid human decision-makers to promptly act, the same cannot be said for the use of AI to screen potential asylum seekers and refugees upon arrival to their host country. Minimal due process guarantees at entry points cannot not be upheld by ADM as discussed below. In short, algorithms are vulnerable to the same decision-making concerns that plague human decision-makers: transparency, accountability, discrimination,

²¹ Patrick Breyer, 'EU-Funded Technology Violates Fundamental Rights' (*About:intel*, 22 April 2021) <<http://aboutintel.eu/transparency-lawsuit-iborderctrl/>>; Umberto Bacchi, 'EU's Lie-Detecting Virtual Border Guards Face Court Scrutiny' (*Reuters*, 5 February 2021) <www.reuters.com/article/europe-tech-court-idUSKBN2GT>.

²² Molnar (n 18).

²³ Shane Ferro, 'ICE's Bond Algorithm has One Response: Detain' (*Above The Law*, 27 June 2018) <<https://abovethelaw.com/2018/06/ices-bond-algorithm-has-one-response-detain/>>.

²⁴ IACtHR (n 4) para 163; *Pacheco Tineo Family v Bolivia*, Preliminary Objections, Merits, Reparations and Costs, Judgment, Inter-American Court of Human Rights Series C No 272 (25 November 2013); *Othman (Abu Qatada) v the United Kingdom* App no 8139/09 (EctHR, 17 January 2012); UNHCR, 'Guidelines on the Applicable Criteria and Standards relating to the Detention of Asylum-Seekers and Alternatives to Detention' (UNHCR, 2012); IACtHR, 'Inter-American Principles on the Human Rights of All Migrants, Refugees, Stateless Persons, and Victims of Human Trafficking' OAS Res 04/19 (2019).

²⁵ IACtHR (n 4) para 228.

bias, and error.²⁶ The use of these AI technologies without norms to regulate them raises serious concerns not only about the right to seek asylum and to have access to the territory to do so, but also regarding the right to privacy of non-citizens who have not consented to give their data to these algorithms.

2.2 The Right to File for Asylum

Being able to file for asylum and registering as an asylum seeker is a core component of the right to seek asylum. Thus, registration by states or agencies involved in forced migration management, like the UNHCR, is a key first step in ensuring their protection. For the UNHCR, together with registration, the use of biometrics provides an accurate way to verify identities using unique physiological characteristics, such as fingerprints, irises, and facial features, which ensure that refugees' personal identities cannot be lost, registered multiple times, or subject to fraud or identity theft.²⁷ However, the collection of biometrical data is of grave concern and could pose life-threatening risks for individuals forced to flee if this data is mishandled, whether through data sharing agreements, leaks, or criminal hacking.²⁸

Besides their registration uses, the UN has implemented humanitarian aid distribution programmes based on biometric data alone. For example, both the UNHCR and the World Food Programme (WFP) use iris-recognition software for the delivery of humanitarian assistance. The UNHCR has implemented iris-scanning technologies for asylum seekers and refugees to receive a direct cash allowance. And, in partnership with the WFP, the UNHCR has implemented an iris scan and electronic voucher programme for Syrian refugees in Jordan.²⁹

Even though consent is required to collect individuals' biometric data, research shows that this consent is often biased and at times coerced.³⁰ If individuals were to refuse the collection of their biometric data, while in theory this should not affect their rights, in practice individuals could be cut off from the registration process, which in turn leads to the denial of benefits such as cash-based intervention, food

²⁶ Petra Molnar and Lex Gill, 'Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System' (Research Report No 114, University of Toronto, September 2018).

²⁷ UNHCR, 'Guidance on Registration and Identity Management' (2020) <www.unhcr.org/registration-guidance/>.

²⁸ Elise Thomas, 'Tagged, Tracked and in Danger: How the Rohingya got Caught in the UN's Risky Biometric Database' (*Wired Magazine*, 12 March 2018) <www.wired.co.uk/article/united-nations-refugees-biometric-database-rohingya-myanmar-bangladesh>.

²⁹ Leah Waid, 'Tracing the Untraceable: New Technologies that Monitor Displaced Persons and the History of Population Control' (*Harvard International Review*, 24 May 2021) <<https://hir.harvard.edu/new-technologies-that-monitor-displaced-persons>>; UN News, 'Iris Scan Helps Syrian Refugees in Jordan Receive UN Supplies in "Blink of Eye"' (UN News, 6 October 2016) <<http://news.un.org/en/story/2016/10/542032-iris-scan-helps-syrian-refugees-jordan-receive-un-supplies-blink-eye>>.

³⁰ Molnar (n 6) 313; Waid (n 29).

assistance, and other humanitarian interventions. Informed and free consent to collect biometric data is hard to obtain from the population of concern when there are power dynamics at play, language and cultural barriers, fear of denial of benefits, and without a true understanding how their data will be collected, used, and shared.

The biggest concern with biometric data is who can access it, particularly since once collected or shared, such biometric data is practically impossible to discard. For example, in June 2021, a Human Rights Watch (HRW) report revealed that the UNHCR improperly collected and shared personal information from ethnic Rohingya refugees with the Government of Bangladesh, which, in turn, shared it with Myanmar to verify people for possible repatriation, even when returning Rohingya refugees to Myanmar would put them at grave risk of arbitrary arrest, torture, ill-treatment, and possible death.³¹ This is not a new occurrence, history provides many examples of the discriminatory and even deadly use of data collection from marginalised groups, from Nazi Germany collecting information on Jewish people to the Tutsi Genocide in Rwanda.³²

While biometric data is not by itself an AI, when used in combination with other automated technologies to identify and make decisions regarding asylum seekers and refugees, it can directly affect their rights. It is of particular concern when this data is improperly used or shared to deny people access to territory or humanitarian aid or to return people to where their lives would be at risk.³³

2.3 Refugee Status Determination Process

The 1951 UN Convention Relating to the Status of Refugees (Refugee Convention) enshrines all the rights that refugees are entitled to. In order for persons forced to flee to access said rights, they must be deemed to meet the refugee definition. For the Refugee Convention, a refugee is ‘any person who owing to well-founded fear of persecution for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of their nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country’.³⁴ Even though the refugee status is declarative and not constitutive, it is still a threshold requirement for the attainment of specific rights and entitlements set out in the Refugee Convention and/or national legislation.

³¹ Human Rights Watch (HRW), ‘UN Shared Rohingya Data Without Informed Consent’ (15 June 2021) <www.hrw.org/news/2021/06/15/un-shared-rohingya-data-without-informed-consent>; Thomas (n 28).

³² UNGA (n 5).

³³ Kinchin (n 9); Molnar (n 6) 313.

³⁴ Refugee Convention, art 1.A(2).

The process by which a competent authority (whether it be a state or the UNHCR) confers refugee status to an individual that has filed for asylum is called ‘Refugee Status Determination’ (RSD). RSD is an administrative procedure with many due process guarantees to ensure asylum seekers have a fair and efficient procedure in the evaluation of their claim. While all of the guarantees are essential to uphold the right to seek and enjoy asylum, AI such as ADM could affect core areas of this process, particularly:

- (i) the right to a personal interview, whereby the person can explain the individualised circumstances of their case;
- (ii) the right to a reasoned and substantiated decision on the determination of their refugee status; and
- (iii) the right to appeal said decision, of particular importance if the claim has been rejected.

There are compelling reasons to look at the use of AI to support the processing of large caseloads of asylum applications, particularly when many asylum systems—including the UNHCR which conducts RSD under their mandate—are plagued by lengthy delays in their RSD processing which directly affect asylum seekers. AI facilitates faster data processing and the ability to undertake a higher volume of repetitive tasks and can ultimately lead to ADM processes.

For the purposes of RSD, ADM can take various forms but, ultimately, its goal is to assist or replace the judgment of human decision-makers in the RSD process. AI could assist in determining the priority for the case such as, researching the country of origin information (COI) to support or refute the applicant’s claim, reviewing and assessing submitted evidence, and could autonomously make a decision on whether an applicant should be granted refugee status or not, without human intervention. AI technologies are extremely high-risk within asylum systems that are already plagued with arbitrary and opaque decision-making, along with ineffective remedies. These risks are especially high when the result of a negative outcome can result in the death, detention, or harm, of the rejected asylum seeker.

There are multiple concerns about these algorithms that would support ADM. First, where does the ‘input data’ come from and how was it collected? Notwithstanding widespread perceptions of emerging digital technologies as neutral and objective in their implementation, it is inescapable that race, ethnicity, national origin, and citizenship play a large role when these technologies are implemented.³⁵ ADM algorithms assume the future will look like the past, and when the past is unfair or biased, machine learning (ML) will propagate these biases and

³⁵ UNGA (n 5).

enhance them through feedback loops.³⁶ These types of discriminatory outputs can be difficult to correct—or even to detect in the first place—particularly when the biased result comes from the presumed neutral ‘input data’.³⁷

A second concern is the uninterpretable and unexplainable conclusions the algorithms will present as these technologies begin to learn, iterate, and improve upon themselves. RSD decisions are already highly discretionary, so much so that a case may be analysed by two different adjudicators who could arrive at different conclusions. AI for these cases would mean relying on machines that derive conclusions from models that they themselves have created, and that cannot be understood or explained by a human decision-maker.³⁸

Regional human rights courts have already interpreted the obligation to provide a reasoned and substantiated decision within asylum adjudicating procedures to mean that ‘the decision on the request taken by the competent authority as to whether the applicant is granted refugee status based on the factual and legal determinations must expressly include the reasons for the decision, in order to enable the applicant to exercise their right of appeal’.³⁹

The use of AI poses challenges to this due process guarantee, given an algorithm’s inability to explain itself or verbalise reasons for its decisions the way that a human decision-maker would be able—and required—to do. The lack of transparency in algorithms may constitute a denial of procedural fairness.⁴⁰ An automated decision would also directly affect the individual’s right to appeal negative decisions, as the applicant would be unable to understand—much less appeal—the grounds for the denial.

A positive safeguard for the use of automated decisions is that no final negative decision should be made by an AI and all recommendations made by such systems need to be reviewed by a human decision-maker. For example, Canada’s Department of Immigration, Refugees, and Citizenship (IRCC) has resorted to the use of ‘advanced data analytics’ to sort and process temporary resident visa applications. In particular, the AI reviews initial applications, sorts and triages them, but is only able to make recommendations in low complexity cases that will be approved. The final decision on all applications, whether to approve or deny, would be made by an IRCC officer.⁴¹

Generalised delays and backlogs in asylum systems ultimately have a negative impact on asylum seekers who have been waiting for years to be recognised

³⁶ Molnar and Gill (n 26) 9.

³⁷ ibid 33.

³⁸ ibid 11.

³⁹ *Rights and Guarantees of Children in the Context of Migration and/or in Need of International Protection*, Advisory Opinion OC-21/14 (IACtHR, 19 August 2014), para 257.

⁴⁰ Molnar and Gill (n 26) 62.

⁴¹ IRCC, ‘Advanced Data Analytics to Help IRCC Officers Sort and Process Temporary Resident Visa Applications’ (24 January 2022) <www.canada.ca/en/immigration-refugees-citizenship/news/notices-analytics-help-process-trv-applications.html>.

as refugees and be able to exercise their rights; however, the use of AI is not the sole answer to this problem. Certainly, when properly regulated, AI applications that assist in the reduction of clerical and repetitive tasks and make the RSD process more efficient, present an advantage for asylum authorities and have a direct positive impact on asylum seekers. Yet, when these algorithms make the wrong decision, the consequences can be grave and hard to unveil. For example, in the United Kingdom (UK), approximately 7,000 students were deported because an algorithm wrongly accused them of cheating on a language test, after using voice recognition software to allegedly identify fraudulent test takers.⁴² In short, states would be unable to fulfil their international obligations if human decision-makers are replaced by algorithms that are inherently opaque, biased, and unfair.

2.4 Durable Solutions that End the Displacement Cycle

Being recognised as a refugee does not necessarily put an end to the displacement cycle of people forced to flee. Refugee status was envisioned as a temporary status to provide protection until these individuals could find a durable solution. The UNHCR defines a durable solution as any means by which the situation of refugees can be satisfactorily and permanently resolved to enable them to live normal lives; these are traditionally considered to be: voluntary repatriation, local integration, or resettlement.⁴³

The current uses of AI for durable solutions have focused on the placement of resettled refugees. Resettlement is the procedure by which a third country receives a refugee who has fled their country of origin and resides in a country of asylum where their rights and needs are not met. Resettlement countries agree to admit these refugees, grant them permanent residence status, and the opportunity to naturalise; traditional resettlement countries are those of the Global North. When refugees are resettled to those countries, local non-governmental organisations (NGO) are in charge of placing them and assisting them with integration.

For example, an AI-powered software called 'Annie MOORE' uses complex computational tools to match refugees to a given location by their needs, skills, and the number of available resources and opportunities available. The software is able to provide recommendations on simple cases, according to the criteria that the local NGOs have set. There has been positive feedback with the use of this software as it saves time on straightforward placements and allows time to properly address complex cases. The software's recommendation is not final, as it must be confirmed

⁴² Ed Main and Richard Watson, 'The English Test That Ruined Thousands of Lives' (BBC News, 9 February 2022) <www.bbc.com/news/uk-60264106>.

⁴³ UNHCR, 'UNHCR Master Glossary of Terms' (2021) D <www.unhcr.org/glossary/>.

or rejected by a human caseworker.⁴⁴ One study has concluded that the use of such software in the past would have increased the average employment rate of refugees upon resettlement.⁴⁵

While the use of these automated systems seems promising in processing large amounts of cases in a timely manner and reducing administrative costs, the software can replicate the same bias and lack of transparency previously mentioned, generating discriminatory outputs. These technologies focus only on the likelihood of someone becoming employed, disregarding other needs these refugees might have or the communities they will be placed in. Systems like these ones can reinforce and exacerbate inequalities by placing refugees with the least prospect of success into under-resourced areas, thus perpetuating cycles of poverty.⁴⁶

Refugee placement in resettlement processing is a key component to ensure the newly arrived population can have a successful social, economic, and legal integration that upholds their rights. Otherwise, those refugees who have already survived violence and persecution could find themselves in situations replicating the same issues that led many of them to be resettled in the first place.

3 Regulations of AI Technologies Used in Forced Migration Management

Current international refugee law instruments do not directly address the use of AI in forced migration management. Since 2017, at least sixty countries have adopted some form of AI policy.⁴⁷ However, there are no international treaties to regulate AI technologies in the forced-migration context, facilitate positive effects, or mitigate risks. Still, the regulatory international human right framework provides a strong basis for obligations and standards that should be at the forefront of all responses to the new developing technologies.

There are some regional efforts that touch on AI that could impact the right to seek and enjoy asylum. For example, article 22 of the EU's General Data Protection Regulation (GDPR) addresses 'automated individual decision-making, including profiling'.⁴⁸ The regulation mandates consent of the individual to automated

⁴⁴ University of Oxford, 'Using AI to Improve Refugee Integration' (2 October 2018) <www.ox.ac.uk/news/2018-10-02-using-ai-improve-refugee-integration>.

⁴⁵ Kirk Bansak and others, 'Improving Refugee Integration through Data-Driven Algorithmic Assignment' (2018) 359 *Science* 325.

⁴⁶ Molnar and Gill (n 26) 39.

⁴⁷ Alex Engler, 'The EU and US Are Starting to Align on AI Regulation' (*Brookings*, 1 February 2022) <www.brookings.edu/blog/techtank/2022/02/01/the-eu-and-u-s-are-starting-to-align-on-ai-regulation/>.

⁴⁸ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (General Data Protection Regulation/GDPR), art 22.

individual decision-making, the right not to be subject to a decision based solely on automated processing, and the right to obtain human intervention. The GDPR is not only applicable to EU countries, but to any company or government that may have access to the files of an EU citizen.

In April 2021, the European Commission submitted its proposal for a EU regulatory framework on AI. The Artificial Intelligence Act represents the first attempt globally to regulate AI. This regulation is of high relevance as it will probably serve as a template for other regional, domestic, or international regulations on the matter. The draft regulation has been welcomed by experts in the field as a positive step forward, as it classifies as high-risk:

AI systems intended to be used by the competent public authorities charged with tasks in the fields of *migration, asylum and border control management* as polygraphs and similar tools or to detect the emotional state of a natural person; for assessing certain risks posed by natural persons entering the territory of a Member State or applying for visa or asylum; for verifying the authenticity of the relevant documents of natural persons; for assisting competent public authorities for the examination of applications for asylum, visa and residence permits and associated complaints with regard to the objective to establish the eligibility of the natural persons applying for a status.⁴⁹

The regulation indicates that AI systems used in migration, asylum, and border control management affect people who are often in particularly vulnerable position and who are dependent on the outcome of the actions of the competent public authorities. Thus, the accuracy, non-discriminatory nature, and transparency of the AI systems used in those contexts are particularly important to guarantee the respect of the fundamental rights of the affected persons, notably their rights to free movement, non-discrimination, protection of private life and personal data, international protection, and good administration. AI systems classified as high-risk will have to comply with a set of horizontal mandatory requirements for trustworthy AI and follow conformity assessment procedures before those systems can be placed on the EU market.

Moreover, international organisations such as the UNHCR and the International Organization for Migration (IOM) pose an additional challenge for the regulation of AI and the guaranteeing of rights, since UN Agencies cannot compromise their international responsibility. Left to their own internal regulations and without accountability, the use of AI by UN agencies involved in migration management is as

⁴⁹ European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD), Recital 39 (emphasis added) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

much a threat to the rights of people forced to flee as the use of these technologies by states and the private sector.⁵⁰

4 Conclusion

People forced to flee have to traverse a myriad of hurdles to arrive at safety, from fleeing their oppressive governments, travelling through dangerous terrains, and ultimately arriving at a country of asylum, where they then need to prove they need international protection and cannot be returned home. The right to seek and enjoy asylum is a wide-ranging prerogative because it protects those that have already suffered persecution and multiple human rights abuses, and mandates other countries to grant international protection precisely because domestic protection is unavailable or unattainable. The journey of forced migration is far from simple, and so are the individual stories of the millions of people forced to flee their homes. As leading scholar in the field, Petra Molnar has stated that ‘the complexity of forced migration cannot be reduced to an algorithm’⁵¹

Even with well intentioned policies to improve the management of large caseloads, AI technologies are not an easy fix for asylum systems that are already highly discretionary and do not fulfil international obligations under the right to seek asylum. Efficiency and cost reduction are certainly important factors. However, turning to biased and opaque AI algorithms that can track, monitor, and make life-or-death decisions on asylum cases not only directly breaches many of the obligations deriving from the right to seek asylum, but it also encourages discriminatory practices, increases risks to life and liberty, and leads to the denial of due process and procedural safeguards. While AI can certainly have a positive effect when properly regulated and used to assist human decision-makers on non-substantial case management aspects, it is important to limit their use to comply with international human rights standards and obligations.

The international community has an obligation to consider not only the ‘well-intended’ uses of these technologies. Once developed, the very same AI algorithms that claim to save lives by detecting people at high-sea or deserts, distributing humanitarian assistance through iris scans, and making automated decisions without human intervention, could be the same technologies used to turn away people seeking international protection, to segregate populations, to automate mass rejection decisions, and deport people back to countries where their lives and liberties are at risk.

UN experts, academia, think tanks, and other stakeholders have already amassed a series of specific recommendations for the use of AI technologies in procedures

⁵⁰ Molnar (n 18) 141.

⁵¹ Molnar and Gill (n 26) 34.

related to asylum seekers, refugees, and others forced to flee.⁵² Their main recommendations for states, international organisations, and the private sector relate to:

- (i) placing an immediate moratorium on the procurement, sale, transfer, and use of surveillance technology and automated migration management technologies, until robust human rights safeguards are in place to regulate such practices, since it is essential that the regulatory framework for AI complies with international human rights law, independent oversight, strict privacy and data protection laws, and full transparency;
- (ii) the need to establish an independent body to oversee and review all use of existing and proposed automated technologies in migration management; and
- (iii) outright banning technology in migration management that are high-risk and that cannot meet the standards enshrined in international human rights legal frameworks.

If such technologies are deployed to manage forced migration, recommendations are aimed at the need to:

- (i) establish guarantees that no decision which prejudicially impacts the rights of people forced to flee is made by an automated decision system alone;
- (ii) require express consent from individuals to use an automated decision system in their case and guarantee the right to obtain human intervention;
- (iii) ensure automated systems are designed to err on the side of outcomes with the lowest impact on rights; and
- (iv) provide for clear processes to facilitate complaint, review, redress, and appeal of decisions made by automated systems.

A final cross-cutting recommendation is the need to ensure transparency and accountability for private and public sector, as well as international organisations, for the use of AI in migration management by enabling independent analysis and oversight, adopting mandatory human rights impacts assessments, and providing persons forced to flee with mechanisms for direct accountability for violations of their human rights resulting from the use of these technologies.

At the centre of the use of these AI technologies in migration management is the reality that these systems are used on non-citizens to track, identify, and control them, particularly because migrants, refugees, and asylum seekers are not able to exercise the same rights as citizens. Technology itself replicates existing power

⁵² UNGA (n 5); see also Petra Molnar, ‘Technological Testing Grounds: Border Tech Is Experimenting with People’s Lives’ (*EDRI*, November 2020) <<https://edri.org/wp-content/uploads/2020/11/Technological-Testing-Grounds.pdf>>; Molnar and Gill (n 26) 63; Molnar (n 6).

hierarchies even within voluntary and forced migration. Without a fundamental shift away from racist, xenophobic, anti-refugee, and anti-migrant narratives and political approaches to border governance, the discriminatory effects of AI highlighted above cannot be redressed.⁵³ In conclusion, the nuanced and complex nature of forced displacement cannot be properly addressed by these new emergency technologies without breaching international obligations on the right to seek and enjoy asylum from persecution.

⁵³ UNGA (n 5).

22

Artificial Intelligence Screening and the Right to Asylum

Dhruv Somayajula

1 Introduction

Throughout world history, there are numerous instances of refugees fleeing their places of origin to escape violence or persecution. Recent world events such as war, persecution, and sectarian violence have resulted in a large-scale migration of refugees fleeing Somalia, Syria, Afghanistan, and Ukraine.¹ The right to asylum, afforded to any individual seeking asylum from persecution, is a human right enshrined in the 1948 Universal Declaration of Human Rights (UDHR)² and further set out in various instruments of international law. However, a refugee's eligibility to seek asylum in another country as a refugee depends on the persecution from which they are fleeing. Evaluating this eligibility requires undertaking a complex set of functions—interviewing the applicant while independently verifying the persecution under consideration, whilst, moreover, delicately handling human vulnerability and suffering. Under international refugee law, this process of refugee status determination (RSD) assumes great importance, with the stakes in question involving a human right and an offer to a safe sanctuary for persons in need. In the twenty-first century, artificial intelligence (AI) has rapidly been adopted as a tool to augment human decision-making capabilities. AI today plays a vital role across various sectors, including transport,³ healthcare,⁴ and education.⁵ The rise in AI

¹ Boris Cheshirkov, 'UNHCR Ramps Up Aid to Thousands Displaced by Somalia Drought' (*UNHCR Briefing Notes*, 11 March 2022) <www.unhcr.org/news/briefing/2022/3/622b03ba4/unhcr-ramps-aid-thousands-displaced-somalia-drought.html>; Charlie Dunmore, 'The Refugee Brief: 18 March 2022' (*UNHCR The Refugee Brief*, 18 March 2022) <www.unhcr.org/refugeebrief/the-refugee-brief-18-march-2022/>.

² Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR), art 14.

³ European Parliamentary Research Service, 'Artificial Intelligence in Transport' (March 2019) <[www.europarl.europa.eu/RegData/etudes/BRIE/2019/635609/EPRI_BRI\(2019\)635609_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/BRIE/2019/635609/EPRI_BRI(2019)635609_EN.pdf)>.

⁴ Sandeep Reddy, John Fox, and Maulik Purohit, 'Artificial Intelligence-Enabled Healthcare Delivery' (2019) 112(1) *Journal of the Royal Society of Medicine* 22.

⁵ Xiaozhe Yang, 'Accelerated Move for AI Education in China' (2019) 2(3) *ECNU Review of Education* 347.

development and implementation has corresponded with an increased appetite for reducing the human element through AI-based automation in public services such as human decision-making and financial risk identification roles across the world.⁶ Perhaps naturally, the potential benefits of AI systems in RSD processes and in the broader humanitarian context are currently being examined across various countries. In the past few years, AI systems have been deployed in recommendatory screening systems, predictive forecasting, migration management, and credibility testing of refugees. The right to asylum, guaranteed under international humanitarian law, hangs in the balance in the operations of these AI-based systems. AI systems present unique considerations which may affect this right. Therefore, it is expedient to discuss the potential concerns unique to the use of AI systems in the humanitarian context.

The chapter is structured as follows:

- (i) Section 2 describes the obligations of states arising from the right to asylum and the right against discrimination under international law.
- (ii) Section 3 details the use of AI systems and its risks in the refugee context globally.
- (iii) Section 4 proposes certain policy changes to address the concerns raised by the use of AI systems detailed in this chapter.

2 The Right to Asylum under International Law

This section seeks to discuss the rights of refugees under international law, focusing on the right to asylum and the right against discrimination. These legal principles gain further relevance considering the use of AI systems for screening refugees at border checkpoints, as is explained in section 3. Rights accorded to refugees, such as the right to asylum and principle of non-refoulement, are a result of treaty obligations applicable to various countries. Treaty obligations comprise the bulk of international law jurisprudence which form the basis for humanitarian laws in signatory countries.

A discussion on the use of AI in the refugee context is better framed by first summarising the key legal obligations relating to refugees under international law. Refugees are protected through international human rights laws and international humanitarian laws, through a network of international treaty obligations and domestic laws adopted by signatories. The 1951 Convention Relating to the Status of Refugees, as amended by the 1967 Protocol Relating to the Status of Refugees

⁶ Adamantia Rachovitsa and Niclas Johann, ‘The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch *SyRI* Case’ (2022) 22 Human Rights Law Review 1.

(collectively, 'the Refugee Convention') specifically sets out rights and obligations owed to refugees and asylum applicants. Additionally, treaty obligations under the UDHR and 1966 International Covenant on Civil and Political Rights (ICCPR) set out the right to asylum, which together with the right to equality and non-discrimination, is relevant for AI applications that may provide inaccurate outputs and result in a violation of these rights.

The right to seek and enjoy asylum from persecution is part of the UDHR.⁷ This right has further been expanded through the Refugee Convention. Under the Refugee Convention, contracting states are required to apply its provisions without discrimination on race, religion, or country of origin.⁸ Moreover, any expulsion of a refugee shall only be made pursuant to a decision reached under due process of law. The Refugee Convention clarifies that 'due process' includes obligations to give a refugee the right to be heard and present evidence to clear themselves before the competent authority designated to hear such disputes.⁹

The Refugee Convention prohibits the refoulement, or turning back, of any refugee to the territories where their life or freedom would be threatened on grounds that include race, religion, or nationality.¹⁰ Under the Refugee Convention, the idea of refoulement includes rejecting a refugee at the border from admission into the country, and is accordingly prohibited as well.¹¹ This principle has been unanimously adopted by the United Nations General Assembly (UNGA) in its 1967 Declaration on Territorial Asylum, which expressly provides that persons seeking asylum under article 14 of the UDHR shall not be subject to rejections at the frontier.¹² The principle of non-refoulement has since been accepted as a *jus cogens* norm, forming a part of customary international law, and is applicable to countries which have not acceded to the Refugee Convention as well.¹³

There are several non-discrimination protections available under international human rights law for refugees with respect to state agencies and their immigration officials using AI systems as an extension of its functions. Article 26 of the ICCPR entitles individuals to equal protection of the law without any discrimination on race, colour, sex, or religion.¹⁴ The right to non-discrimination is also provided in the context of other international treaty obligations.¹⁵ Discrimination and its

⁷ UDHR, art 14(1).

⁸ Convention Relating to the Status of Refugees (adopted 28 July 1951, entered into force 22 April 1954) 189 UNTS 137 (Refugee Convention), art 3.

⁹ *ibid* art 32(2).

¹⁰ *ibid* art 33.

¹¹ P Weis, *The Refugee Convention, 1951: The Travaux Préparatoires Analysed with a Commentary by Dr Paul Weis* (1st edn, CUP 1995) 244, 272.

¹² Declaration on Territorial Asylum, UNGA Res 2312(XXII) (adopted 14 December 1967), art 3(1).

¹³ Jean Allain, 'The Jus Cogens Nature of Non-Refoulement' (2001) 13(4) International Journal of Refugee Law 533.

¹⁴ International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR), art 26.

¹⁵ International Convention on the Elimination of All Forms of Racial Discrimination (adopted 21 December 1965, entered into force 4 January 1969) 660 UNTS 195 (ICERD); Convention on the

consequences upon refugees due to inaccurate decisions based on bias by AI systems needs further review against the legal principles and safeguards provided on paper in signatory countries.

In accordance with UNGA resolutions, countries may rely on expedited entry systems at border checkpoints in cases where claims are clearly abusive or manifestly unfounded.¹⁶ In such instances, the state agencies using expedited entry systems are required to provide certain safeguards in line with the consequences of an erroneous determination for a refugee. These safeguards include conducting complete personal interviews with all refugee applicants, permitting only a competent authority capable of determining refugee status to decide on whether a claim is clearly abusive or manifestly unfounded, and granting an unsuccessful interviewee applicant with a chance to review a negative decision prior to being returned from a border or removed from the country of refuge.¹⁷ The use of AI systems at the border such as facial recognition, voice analysis, and truth detection can be deemed to be expedited entry systems, and the foregoing safeguards must be applied to the use of these AI systems concurrently.

3 AI Systems that Impact the Right to Asylum

This section seeks to explain the purposes and types of AI systems being used in the refugee context. Section 3.1 explains the need for using AI systems in this sector. Examples of specific use cases are detailed in section 3.2, which include predictive analytics, credibility testing, recommendatory decision-making systems, and migration management tools that provide various levels of support to existing immigration and refugee systems across the world.

3.1 Why Are AI Systems Being Used for Refugees?

In the past few years, refugee applications have exponentially increased across the world, with refugees fleeing war zones, climate crises, and sectarian or

Elimination of All Forms of Discrimination Against Women (adopted 18 December 1979, entered into force 3 September 1981) 1249 UNTS 13 (CEDAW); Convention on the Rights of the Child (20 November 1989, entered into force 2 September 1990) 1577 UNTS 3 (CRC); Convention on the Rights of Persons with Disabilities (adopted 13 December 2006, entered into force 3 May 2008) 2515 UNTS 3 (CRPD).

¹⁶ Executive Committee of the High Commissioner's Programme, 'The Problem of Manifestly Unfounded or Abusive Applications for Refugee Status or Asylum' (20 October 1983) No 30 (XXXIV); see also Committee of Ministers, 'Guidelines on Human Rights Protection in the Context of Accelerated Asylum Procedures' (1 July 2009) Council of Europe <www.refworld.org/docid/4a857e692.html>.

¹⁷ Executive Committee of the High Commissioner's Programme, 'The Problem of Manifestly Unfounded or Abusive Applications for Refugee Status or Asylum' (20 October 1983) No 30 (XXXIV).

gender-based violence. By mid-2022, an estimated 4.9 million persons have submitted applications seeking asylum.¹⁸ This efflux of refugees has strained the resources of traditional RSD methods.

On an individual basis, many countries have faced manpower constraints in terms of RSD officials and resettlement programmes. Asylum applicants may be forced to wait for over four years to resolve their cases in the United States (US),¹⁹ while the average wait time in the United Kingdom (UK) is reportedly between one and three years.²⁰ This increasing backlog and time taken during pendency of asylum applications has led to the implementation of AI systems in border checkpoints to expedite entry and settlement for refugees. It has also forced novel approaches to predicting refugee crises and processing applications seeking asylum.

Today, AI systems are being adopted for assistance across humanitarian concerns. The use-cases of AI in the asylum context includes the use of AI-based systems for predictive analytics used to predict international refugee trends, AI systems relying on processing biometric data or voice analysis to cross-reference the veracity of asylum applications, recommendatory automated decision-making (ADM) systems used to score or flag refugees based on specified criteria, and algorithmic ‘migration management’ tools used for optimally allocating refugees to specific areas within a country granting asylum.

3.2 The Use of AI Systems for Refugees

The use of AI systems has predominantly been observed in the context of refugee forecasting, migration management, and screening of refugees. A major application of AI systems on refugees is its use as a screening tool. AI can be used to screen faces for authentication, voices to test credibility, and to screen asylum applications and recommend a decision to the human-in-charge. Systems that track domestic and international displacement are relevant for countries that may need to build capacity accordingly, while migration management tools aim to allocate refugees based on specific regions to maximise their employment and social opportunities. This section also highlights the use of AI systems in preparing credibility scores and providing border officials with recommendations on asylum applications. These use-cases highlight the current prevalence of AI systems and underline the urgent need to mitigate potential AI risks impacting refugee rights.

¹⁸ United Nations High Commissioner for Refugees, ‘Key Indicators-Refugee Data Finder’ (27 October 2022) <<https://www.unhcr.org/refugee-statistics/>>.

¹⁹ TRAC Immigration, ‘A Mounting Asylum Backlog and Growing Wait Times’ (22 December 2021) <<https://trac.syr.edu/immigration/reports/672/>>.

²⁰ Refugee Council, ‘Thousands Seeking Asylum Face Cruel Wait of Years for Asylum Decision: Fresh Research Shows’ (Refugee Council, 2 July 2021) <<https://refugeecouncil.org.uk/latest/news/thousands-seeking-asylum-face-cruel-wait-of-years-for-asylum-decision-fresh-research-shows>>.

3.2.1 Predictive Analytics

In 2020, the European Commission published a study in which the use of AI-based tools for migration forecasting were discussed.²¹ This system was deemed necessary in predicting irregular migratory trends, which would allow EU member states to assess and build refugee capacity as well as bolster border management procedures. Similar AI-based models are currently being trained to identify trends in forced migration within Africa.²² Furthermore, AI solutions are also aimed at monitoring individuals crossing the border between Venezuela and Brazil, by modelling simulated border crossings, estimating current urban populations, and predicting future arrivals.²³ This tool has helped UNHCR Brazil to prepare medical service and shelter for arrivals, including isolation areas to treat COVID-19 cases. These algorithmic models are also feasible due to the presence of digital tools that track and monitor displaced individuals such as the Displacement Tracking Matrix.²⁴ Additionally, the UNHCR has launched a machine learning (ML) algorithmic model named 'Project Jetson', currently used to predict the movement of internally displaced persons within Somalia.²⁵ The adoption of such software programs indicates the growing relevance of predictive refugee forecasting at a global level.

3.2.2 Automated Credibility Tests

Recent developments in AI include the ability to use AI to detect deception from answers given by individuals crossing the border.²⁶ The Federal Office for Migration and Refugees in Germany (BAMF) has deployed AI-based facial recognition to evaluate the credibility of asylum applicants.²⁷ Furthermore, the BAMF uses the AI-based 'Dialect Identification Assistance System' which suggests the possible country of origin of Arabic-speaking asylum seekers based on dialect recognition.²⁸ Such AI-based authentication requires processing biometric facial data

²¹ Directorate-General for Migration and Home Affairs, 'Feasibility Study on a Forecasting and Early Warning Tool for Migration Based on Artificial Intelligence Technology: Executive Summary' (European Commission Publications Office, 15 February 2021) <<https://op.europa.eu/en/publication-detail/-/publication/5afa29f0-700a-11eb-9ac9-01aa75ed71a1/language-en>>.

²² Babusi Nyoni, 'How Artificial Intelligence Can Be Used to Predict Africa's Next Migration Crisis' (UNHCR Innovation Service, 10 February 2017) <www.unhcr.org/innovation/how-artificial-intelligence-can-be-used-to-predict-africas-next-migration-crisis/>.

²³ United Nations Global Pulse, 'Visualizing Venezuela-Brazil Border Scenarios' <<https://brazil-venezuela-flows.unglobalpulse.net/>>.

²⁴ International Organization for Migration, 'Displacement Tracking Matrix' <<https://dtm.iom.int>>.

²⁵ UNHCR Innovation Service, 'Project Jetson' <<https://jetson.unhcr.org/>>.

²⁶ iBorderCtrl, 'Automatic Deception Detection System' <www.iborderctrl.eu/Technical-Framework>.

²⁷ Graeme Wood, 'The Refugee Detectives' (*The Atlantic*, April 2018) <www.theatlantic.com/magazine/archive/2018/04/the-refugee-detectives/554090/>.

²⁸ European Migration Network, 'Ad Hoc Query on 2020.47 Part 1: Procedures for Language Identification by Asylum Authorities' (European Commission, 2021) <https://ec.europa.eu/home-affairs/system/files/2021-01/202047_part_1_procedures_for_language_identification_by_asylumAuthorities.pdf>.

or recorded voice samples to generate a credibility score which is taken into account by RSD officers working on the application. In Hungary, facial recognition technology systems are used by the Aliens Policing Authority to establish the identity of foreign nationals and prevent fraud.²⁹

3.2.3 Recommedatory Automated Decision-Making Algorithms

In recent years, AI systems have been deployed in actively screening applications to enter a country and providing a recommendation to a human official. These recommendations may permit certain individuals to enter the destination country without any hassles. However, applicants that are flagged by such systems may be denied entry by the official involved or may have to undergo additional screening procedures. The current use of these systems in border control checkpoints is relevant to the refugee context. The invasive processing of personal data and bias exhibited in deployment of AI screening systems foreshadow the potential risks of introducing similar systems for refugee intake.

Between 2018 and 2020, Immigration, Refugee and Citizenship Canada (IRCC) has relied on AI and predictive analytics to assist the functions of immigration officials.³⁰ The functions are said to include identifying the merits in an immigration application, detecting potential red flags for fraud, and preparing a recommendation on whether an immigration application is to be approved or rejected by the official. Additionally, the IRCC has launched pilot projects to automate the temporary residence applications from China and India.³¹ This is in addition to the facial recognition technologies being currently tested by Canada under the Known Traveller Digital Identity program.³² This program seeks to build a digital trust score for passengers opting in, with the score calculated based on tracking the interactions of the passenger with banks, hotels, medical and educational institutes, voluntary sharing of credit ratings, educational credentials, vaccination certificates, and travel itineraries.³³ While these pilot programs are currently

²⁹ European Migration Network, 'The Use of Digitalisation and Artificial Intelligence in Migration Management' (EMN-OECD Inform, February 2022) <www.oecd.org/migration/mig/EMN-OECD-INFORM-FEB-2022-The-use-of-Digitalisation-and-AI-in-Migration-Management.pdf>.

³⁰ Nicholas Keung, 'Canadian Immigration Applications Could Soon be Assessed by Computers' (*Toronto Star*, 5 January 2017) <www.thestar.com/news/immigration/2017/01/05/immigration-applications-could-soon-be-assessed-by-computers.html>.

³¹ Teresa Wright, 'Canada's Use of Artificial Intelligence in Immigration Could Lead to Break of Human Rights: Study' (*Global News*, 2018) <<https://globalnews.ca/news/4487724/canada-artificial-intelligence-human-rights/>>.

³² Transport Canada, 'The Government of Canada to Test Cutting-Edge Technologies to Support Secure and Seamless Global Travel for Air Passengers' (Government of Canada, 25 January 2018) <www.canada.ca/en/transport-canada/news/2018/01/the-government_ofcanadatotestcutting-edgetechnologiesupportse.html>.

³³ Tamir Israel, 'Facial Recognition at a Crossroads: Transformation at our Borders and Beyond' (Samuelson-Glushko Canadian Internet Policy and Public Interest Clinic, September 2020) <https://cippic.ca/uploads/FR_Transforming_Borders-OVERVIEW.pdf>.

voluntary, privacy advocates in Canada have warned that widespread adoption of such programs may, at a later stage, cause opting out to become infeasible.³⁴

Similar recommendatory systems have been deployed in the UK. Since 2015, the Home Office of the UK has sought to curb the influx of migrants by deploying an algorithm aimed at streamlining applications submitted by migrants seeking a visa. The algorithm was designed to automatically process visa applications and categorise them into three colour-coded buckets—green, amber, or red—with the colour code rating given by the algorithm playing a crucial role in the treatment and outcome of the application.³⁵ One of the metrics allegedly used to categorise applications into these buckets was nationality, leading to a situation where migrants from certain countries would be flagged under the red bucket, which corresponded with intense scrutiny and a greater probability of rejection. Rejections for applicants from red-flagged countries then was fed into visa decision rates, which identified and informed the algorithm which countries could form part of the red category, creating a feedback-loop which perpetuated an intentionally designed algorithmic bias against certain nationalities.

While nationality is one of the grounds protected from discrimination under the Equality Act 2010,³⁶ enhanced scrutiny of visa applications on grounds of nationality is permitted if notified by a ministerial authorisation.³⁷ To that extent, the classification of certain nationalities under the ‘red’ category followed a legally authorised exercise of intentional bias, with the algorithm classifying the nationalities as it was designed to do.³⁸ The concern, however, arises when a streamlining and categorising function by an algorithm is interpreted as an evaluation of the merits of the application. Reports suggest a higher burden of proof on visa officials that approved a visa application flagged by the AI under the red category, while UK’s Independent Chief Inspector of Borders and Immigration in 2017 pointed out that the streamlining tool was being relied on in a formulaic manner.³⁹

³⁴ ibid.

³⁵ Henry McDonald, ‘AI System for Granting UK Visas is Biased, Rights Groups Claim’ *The Guardian* (29 October 2019) <www.theguardian.com/uk-news/2019/oct/29/ai-system-for-granting-uk-visas-is-biased-rights-groups-claim>

³⁶ Equality Act 2010, s 4 (UK).

³⁷ ibid sch 3.

³⁸ Government of the United Kingdom, ‘Equality Impact Assessment of the Points-Based Immigration System’ <www.gov.uk/government/publications/equality-impact-assessment-of-the-points-based-immigration-system>.

³⁹ David Bolt (Independent Chief Inspector of Borders and Immigration), ‘An Inspection of Entry Clearance Processing Operations in Croydon and Istanbul November 2016–March 2017’ (July 2017) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/631520/An-inspection-of-entry-clearance-processing-operations-in-Croydon-and-Istanbul.pdf>.

3.2.4 Migration Management Tools

Migration management tools provide recommendations for resettlement of a refugee in their destination country. Since 2020, Switzerland has relied on the GeoMatch algorithm⁴⁰ for optimally resettling its accepted refugees into specific provinces. This algorithmic model is trained on past data of asylum seekers and is designed to suggest an optimal location based on likelihood of obtaining employment.⁴¹ This algorithmic recommendation is then reviewed by a human case-worker. Countries such as Canada and the Netherlands have expressed interest in running similar programs.⁴²

3.3 The Risks of Using AI for RSD

The use of AI systems to determine the status of an asylum application raises risks in terms of inaccuracy, bias, opacity, and explainability. RSD officers, under the aegis of the UNHCR, are trained to screen asylum applications, with obligations to empathise and understand refugees,⁴³ and to see each case as a distinct human being.⁴⁴ The use of AI systems for this function may lead to decisions made based on cues derived from a larger, pre-existing data set and risks the system failing to detect unique or newer cases of persecution. This, compounded by the issues of human bias due to automation complacency⁴⁵ or the perception of algorithmic infallibility,⁴⁶ may negatively affect the thoroughness of AI-assisted decisions by RSD officers.

The accuracy of decisions by AI systems in the refugee context may also be affected by error rates⁴⁷ and bias detected within the system. It is necessary to note that the effects of such inaccuracy or bias may prove very costly to refugees seeking

⁴⁰ Immigration Policy Lab, 'Geomatch' <<https://immigrationlab.org/geomatch/>>.

⁴¹ Kirk Bansak and others, 'Improving Refugee Integration through Data-Driven Algorithmic Assignment' (2018) 359(6373) *Science* 325.

⁴² Immigration Policy Lab, 'New Funding Advances GeoMatch in Canada and the Netherlands' (March 2021) <<https://immigrationlab.org/2021/03/03/geomatch-canada-netherlands/>>.

⁴³ UNHCR Code of Conduct 2004, Principle 1 states '*I will always seek to understand the difficult experiences that refugees and other persons of concern to UNHCR have faced and survived, as well as the disadvantaged position in which they – particularly on the basis of gender, age or disability – may find themselves in relation to those who hold power or influence over aspects of their lives.*' See Code of Conduct & Explanatory Notes, UNHCR (June 2004) <<https://www.unhcr.org/media/30248>>.

⁴⁴ ibid para 4.

⁴⁵ Raja Parasuraman and Dietrich Manzey, 'Complacency and Bias in Human Use of Automation: An Attentional Integration' (2010) 52(3) *Human Factors* 381.

⁴⁶ Linda Skitka, Kathleen Mosier, and Mark Burdick, 'Does Automation Bias Decision-Making?' (1999) 51(5) *International Journal of Human-Computer Studies* 991; Sarah Valentine, 'Impoverished Algorithms: Misguided Governments, Flawed Technologies, and Social Control' (2019) 46(2) *Fordham Urban Legal Journal* 364.

⁴⁷ Office of the Information and Privacy Commissioner for British Columbia, 'Investigation into the Use of Facial Recognition Technology by the Insurance Corporation of British Columbia' (BCIPCD No 5, Investigation Report F12-01, 2012), paras 106–12.

entry into a destination country, who may be subject to detention, deportation, or refoulement in instances where persecution is not detected by AI systems. Two major examples of bias in AI systems are *legacy bias*, instituted in its algorithm or based on the training data sets from which it learns its actionable parameters; and *design bias*, actively or passively passed on by the designer of the algorithm. For example, commercial facial recognition algorithms have been shown to match with an error rate of 0.8 per cent for light-skinned men and 38.7 per cent for dark-skinned women.⁴⁸ This raises concerns regarding the lack of heterogeneity within training data sets used to train these algorithms that may contribute to greater inaccuracy, disproportionately affecting refugee applicants from an under-representative facial type. On the other hand, the ‘hostile environment’ policy adopted by the UK displays an example of design bias, where certain countries have been tagged such that entrants from these countries are red-flagged.⁴⁹ Decisions taken on the basis of recommendations provided by algorithmic processes with pre-existing biases could result in grave consequences for refugees fleeing persecution based on their country of origin, ethnicity, gender, sexual identity, and/or sexual orientation.⁵⁰

Another major problem in the use of AI systems in refugee contexts is the problem of accountability. Sufficiently complex AI systems present a challenge in terms of being able to explain the process taken to arrive at a particular decision.⁵¹ These issues materially affect individuals who have received inaccurate or biased outputs as described. Proving bias or exclusion by an uninterpretable AI system requires complainants to have access to technological skills or resources, as well as permission from the developer/operator of the AI system to test its training data sets and source codes on various possible points of bias to prove discriminatory effects.⁵²

The foregoing examples are indicative of wider risks of AI systems being used to process the data of refugees. Any level of harm caused due to discriminatory bias, inaccuracy, and accountability can be exacerbated by the vulnerability of refugees within a third-country’s legal system. Addressing this vulnerability requires policy reforms to foster systems that are transparent, accountable, and minimise the effects of their error rates.

⁴⁸ Larry Hardesty, ‘Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems’ (*MIT News Office*, 11 February 2018) <<https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212>>.

⁴⁹ See section 3.2.3.

⁵⁰ Michael Pizzi and others, ‘AI for Humanitarian Action: Human Rights and Ethics’ (2020) 102(913) International Review of the Red Cross 145.

⁵¹ Cynthia Rudin, ‘Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead’ (2019) 1(5) *Nature Machine Intelligence* 1.

⁵² Yavar Bathaei, ‘The Artificial Intelligence Black Box and the Failure of Intent and Causation’ (2018) 31(2) *Harvard Journal of Law and Technology* 889.

4 Policy Proposal

While the use of AI systems for processing refugees is not directly regulated at the time of writing, relevant regulatory obligations may be sourced from international obligations stemming from the right to asylum and their interaction with national data protection and privacy regulations. The scope of global data protection laws includes the concepts of consent,⁵³ inclusion of personal data in data sets used for training, or processing of personal data through AI systems such as facial recognition.⁵⁴ However, concerns such as opacity, bias, inaccuracy, or interpretability must be addressed within AI systems for their fair, effective, and lawful deployment. For example, national and international legislations seek to ring-fence the use of AI systems by prohibiting any profiling of individuals based solely on the use of AI systems.⁵⁵ This is both an acknowledgment that today's AI systems are, by no means, the finished product and that decisions made by ADM systems have the potential to affect the lives of asylum seekers in both positive and negative ways.

Caution must be exercised in the growing adoption of AI in the refugee context where stakes of inaccuracy or bias may result in harmful consequences on vulnerable individuals. Going forward, immigration and asylum legislation must account for these specific risks while deploying AI systems in varying capacities.⁵⁶ Regulations addressing harms posed by AI systems must be nuanced and granular to account for the variety in the capabilities, purposes, and potential harms of AI systems.⁵⁷ For example, the regulatory approach towards an AI system providing recommendations on asylum applications must be set out differently from the approach taken towards an AI system that uses historical data to undertake migration forecast needs to address capacity building. A specialised ethics agency, set up to examine AI systems used for public functions, would build community trust in the system's reliability and accuracy. The UK's Centre for Data Ethics and Innovation and Singapore's Advisory Council on Ethical Use of AI and Data, set up as independent expert-driven advisory bodies, are examples in this regard.⁵⁸

⁵³ *Commissioner Initiated Investigation into Clearview AI, Inc (Privacy)* 54 AICmr (Office of Australian Information Commissioner, 14 October 2021), para 151.

⁵⁴ *Supreme People's Court Regarding the Trial Use of Face Recognition Technology to Process Personal Information Provisions on Several Issues Concerning the Application of Law in Related Civil Cases* (Supreme People's Court of China, 28 July 2021).

⁵⁵ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (General Data Protection Regulation/GDPR), art 22; Data Protection Act 2018, ss 50, 51 (UK).

⁵⁶ European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD), art 6 <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>

⁵⁷ Mika Viljanen and Henni Parviaainen, 'AI Applications and Regulation: Mapping the Regulatory Strata' (2022) 3 *Frontiers* 1, 6.

⁵⁸ Advisory Board of the Centre for Data Ethics and Innovation (25 April 2022) <www.gov.uk/government/publications/advisory-board-of-the-centre-for-data-ethics-and-innovation>; Infocomm

One key concern highlighted in this chapter has been the inability of a migrant or a refugee to effectively challenge discrimination perpetrated by an ADM system, due to the opaque nature of its source code and training data sets. To this end, certain institutional regulatory measures may be taken up by countries adopting AI systems for refugee processes. Government procurement of AI systems must account for contractual obligations of transparency and accountability.⁵⁹ The accountability stipulations must include obligations to disclose and explain:

- (i) accurate error rates,
- (ii) false positive rates,
- (iii) false negative rates, and
- (iv) the criteria used to obtain this information.⁶⁰

Once operational, AI systems used to assist human decision-making by public authorities must be auditable by a select panel comprising of qualified experts on AI design and ethics.⁶¹ This panel should also include representatives from the countries' national humanitarian bodies that are trained in handling refugee and immigration issues. Such an arrangement can allow a considered approach towards the procurement of reliable and knowledgeable AI systems geared specifically to process refugee data. A periodic assessment of the source code and continuing training sets used to train the AI model, followed up with publicised descriptions of error rates and concerns regarding certain biases if noted, would enable migrants to challenge inaccurate or discriminatory assessments provided by the AI system in immigration tribunals.⁶² Providing the select panel with the output criteria shall increase transparency and enable judicial officers in immigration tribunals to contextualise appeals by affected individuals.

5 Conclusion

This chapter has attempted to explain the diverse use cases of AI systems with respect to individuals seeking asylum, while deliberating on the potential areas of

Media Development Authority, 'Composition of the Advisory Council on the Ethical Use of Artificial Intelligence ("AI") and Data' (26 May 2019) <www.imda.gov.sg/news-and-events/Media-Room/Media-Releases/2018/composition-of-the-advisory-council-on-the-ethical-use-of-ai-and-data>.

⁵⁹ Petra Molnar and Lex Gill, 'Bots at the Gate' (2018) International Human Rights Program and the Citizen Lab 63.

⁶⁰ Israel (n 33).

⁶¹ Molnar and Gill (n 59).

⁶² The use of AI systems for voice analysis was dismissed by an immigration tribunal in the UK, citing the sole use of AI systems as lacking expertise, taking into account its error rate. See *SM and Ihsan Qadir v Secretary of State for the Home Department IA/31380/2014* (Upper Tribunal, Immigration and Asylum Chamber, 21 April 2016).

concern with this deployment and measures to address them. The chapter notes the rise in adoption of AI systems by border officials to identify, authenticate, predict, or score refugees seeking asylum. However, these AI systems raise ethical issues of inaccuracies based on bias or other glitches, as well as the need for transparency and interpretability of AI-enabled decisions.⁶³ The contemporary legal regime around the use of AI systems is not suited to tackle these issues. However, progress is being made in terms of regulating AI systems based on their usage and risks posed. Lastly, the chapter puts out the need to specifically regulate AI systems that play a role in decision-making, which could include AI systems that rank, authenticate, or identify refugees. Legal systems that regulate AI holistically, instead of regulating it as a tangent to data protection laws, may better serve the rights of refugees seeking greater information or grievance redressal. Further, there needs to be greater levels of review and transparency for high-risk AI systems, as well as training officials operating these systems to avoid complacency or automation bias affecting their functions.

⁶³ See the chapter by Kostina Prifti, Alberto Quintavalla, and Jeroen Temperman in this volume.

PART VII

ARTIFICIAL INTELLIGENCE AND SECOND GENERATION RIGHTS

Artificial Intelligence and the Right to Food

Adekemi Omotubora

1 Introduction

The United Nations (UN) has warned that the number of people affected by hunger could surpass 840 million by 2030 despite the guarantee of the right to food under article 11 of the International Covenant on Economic, Social and Cultural Rights (ICESCR) and the commitments of states parties to the realisation of the right.¹ At the 1996 World Food Summit (WFS) (convened by the Food and Agriculture Organization (FAO)) in Rome, heads of states and governments signed the Rome Declaration on World Food Security, thereby making a firm commitment to eradicate hunger and reduce, by 2015, the number of undernourished people by 50 per cent.² They also called for the obligations arising from the right to food under international human rights law (IHRL) to be clarified. However, six years later, at the 2002 WFS, state parties to the ICESCR admitted that the number of undernourished people in the world was falling only by an average of 8 million yearly instead of the target of 22 million people needed to achieve the goals set out in the Rome Declaration.³ There has been no significant change in this number. Instead, hunger and undernourishment have been exacerbated by a chain of events, including natural disasters, population increase, production problems, the COVID-19 pandemic, and the ongoing war in Ukraine.⁴ These events also stunted the progress of UN Sustainable Development Goal (SDG) 2 towards ending hunger, improving nutrition, and promoting sustainable agriculture by 2030. The Food Systems Summit and the Tokyo Nutrition for Growth (N4G) Summit recently proposed

¹ United Nations (UN) Sustainable Development Goal (SDG) 2: ‘Zero Hunger’ <www.un.org/sustainabledevelopment/hunger/>.

² FAO, WFS, ‘Rome Declaration on World Food Security’ (Rome, 13–17 November 1996) <www.fao.org/3/w3613e/w3613e00.htm>.

³ FAO, Declaration of the WFS Five Years Later, para 3<www.fao.org/3/Y7106E/Y7106E09.htm#TopOfPage>.

⁴ See eg European Commission (Knowledge Centre for Global Food and Nutrition Security), ‘The Impact of Russia’s War Against Ukraine on Global Food Security’ (April 2022) <https://knowledge4pol icy.ec.europa.eu/publication/impact-russia-s-war-against-ukraine-global-food-security--kc-fns-review-april-2022_en>.

bold actions, solutions, and strategies—including technological innovations—to put the agenda back on track.⁵

This chapter examines how the ‘smart farm’ can be used to operationalise and realise the right to food. Smart farming involves combined application of ICT, the IoT (Internet of Things), Big Data, sensors, actuators, GPS navigation—and other new technologies such as artificial intelligence (AI)—for a precise and more efficient approach to agriculture.⁶ AI use cases in the ‘smart farm’ are diverse. This makes it suitable for a comprehensive analysis of the positive outcomes of AI across food production and food supply chains (FSCs) on the one hand and, on the other hand, it allows insights into the significant legal challenges of AI in agriculture. This chapter is arranged as follows: section 2 provides a concise discussion of the right to adequate food, examining its scope and core contents, as well as state parties’ obligations under international law. Section 3 examines the smart farm system, its characteristics, and positive outcomes for the right to food. In this section, the chapter analyses how AI in agriculture can create conflicts between the right to food and rights to privacy and decent work and exacerbate food and other systemic inequality within societies and among nations. Section 4 concludes with suggestions on how some of the challenges can be addressed.

2 Meaning and Scope of the Right to Adequate Food

The right to (adequate) food was first recognised in article 25 of the Universal Declaration of Human Rights (UDHR), which provides that everyone has the right to a standard of living adequate for their health and well-being, including food. The ICESCR contains the most explicit articulation of the right. Article 11 recognises ‘the right of everyone to an adequate standard of living for himself and his family, including adequate food, clothing, and housing, and to the continuous improvement of living conditions.’⁷ The ICESCR also recognises the fundamental right of everyone to be free from hunger. It requires states parties to adopt—individually and through international cooperation—the measures, including specific programmes, which are needed:

- (a) to improve methods of production, conservation and distribution of food by making full use of technical and scientific knowledge, by disseminating knowledge of the principles of nutrition and by developing or reforming

⁵ Food Systems Summit Community 2021 <www.un.org/en/food-systems-summit>; Tokyo Nutrition N4G Summit 2021 <https://nutritionforgrowth.org/wp-content/uploads/2021/09/N4G_UN_FoodSysSummit_9.23.pdf>.

⁶ SmartAKIS (Smart Farming Thematic Network), ‘What is Smart Farming?’ <www.smart-akis.com/index.php/network/what-is-smart-farming/>.

⁷ ICESCR, art 11(1).

- agrarian systems in such a way as to achieve the most efficient development and utilisation of natural resources;
- (b) taking into account the problems of both food-importing and food-exporting countries, to ensure an equitable distribution of world food supplies in relation to need.⁸

The most authoritative interpretation of article 11 was provided in General Comment No 12 on the Right to Food (Comment No 12).⁹ The Comment followed a request for legal clarity by states parties to the ICESCR concerning the protection of the right to enhance reporting on obstacles to its realisation.¹⁰ The Comment clarified the scope of the right, its normative content, and the correlative obligations of states. It states that '[t]he right to adequate food is realised when every man, woman, and child, alone or in community with others, have physical and economic access at all times to adequate food or means for its procurement'.¹¹ The concept of adequacy is core to the right and underlines the factors that must be considered in determining whether foods or diets are accessible.¹² Thus, it is not simply the presence of food that must be considered but whether the food is the most appropriate given the circumstances under article 11. For example, the Comment cautions against a narrow interpretation of the right that equates the right to food with a minimum package of calories, proteins, and other specific nutrients. It holds that adequacy covers the quantity and quality of the food (free from adverse substances) and the nutrient (satisfaction of dietary needs), and the non-nutrient (cultural acceptability).¹³ It is also indivisibly linked to the inherent dignity of the human person and indispensable for the fulfilment of other human rights particularly equality and social security, and the right to participate in the cultural life of the community and share in scientific advancement.¹⁴ Food must be accessible in sustainable ways that do not interfere with the enjoyment of other human rights.¹⁵

The core content of the right to adequate food implies availability and accessibility. Food is available when individuals can feed themselves directly from productive land or other natural resources or from effective distribution and market

⁸ *ibid* art 11(2)(a), (b).

⁹ CESCR, 'General Comment No 12: The Right to Adequate Food (Art 11)' adopted at the Twentieth Session of the Committee on Economic, Social and Cultural Rights, on 12 May 1999 (Contained in Document E/C.12/1999/5) (Comment No 12); see also Food and Agriculture Organization (FAO) of the UN, 'Voluntary Guidelines to Support the Realisation of the Right to Adequate Food in the Context of National Food Security' (Adopted by the 127th Session of the FAO Council November 2004) Rome, 2015.

¹⁰ WFS (n 2), Comment No 12, para 2.

¹¹ Comment No 12, para 6.

¹² *ibid* para 7.

¹³ *ibid* paras 7–11.

¹⁴ *ibid* para 4; see also UDHR, arts 22, 23, 27.

¹⁵ Comment No 12, para 9.

systems that can move from production sites to consumers based on demand.¹⁶ In other words, access to land and other resources for food production and an effective FSC is tied to food availability and operationalisation of the right to food. People must have either physical access to resources for production or economic access to purchase food through effective FSCs.

Accessibility encompasses both economic and physical accessibility. Economic accessibility underpins the financial capacity to access food. It means an adequate diet should be affordable for individuals and households or, at least, it should not be attained at a cost detrimental to the attainment of other basic needs. Socially vulnerable groups such as landless persons and other particularly impoverished segments of the population may need attention through special programmes to access food.¹⁷ Physical accessibility implies that adequate food must be accessible to everyone, including those with physical vulnerabilities such as infants, the elderly, disabled, terminally ill, and others who may be particularly vulnerable because of limited access to land, indigenous populations whose access to ancestral land may be threatened, victims of natural disasters and those living in disaster-prone areas.¹⁸ Hence, the normative content containing economic and physical access to food implies the entitlement and access to the means for its procurement, including land and other resources in the widest sense.

States parties to the ICESCR are obligated to *respect, protect, and fulfil* the right to food. The obligation to respect is a negative duty requiring states to refrain from taking measures that prevent access to food. States cannot suspend legislation or policies that provide people with access to food unless fully justified, nor can they arbitrarily evict people from their land, especially if the land was their primary source of subsistence, or knowingly introduce harmful substances into the food chain.¹⁹ The obligation to protect is positive, requiring states to safeguard the right to food against interference by private actors such as individuals or enterprises. States' action to protect may involve the following; one, acting against water pollution or against activities harming the environment by non-state actors and two, enacting consumer protection and food safety laws to prevent food contamination.²⁰ The obligation to *fulfil* incorporates both an obligation to *facilitate* and *provide*. The obligation to *facilitate* means the states must proactively engage in activities intended to strengthen people's access to and utilisation of resources and means to ensure their livelihood, including food security. States have an obligation to *provide* the right direction whenever an individual or group is unable,

¹⁶ ibid para 1.

¹⁷ ibid para 13.

¹⁸ ibid para 13.

¹⁹ See ICESCR, art 11(2); Comment No 12, para 14; see also J Ziegler, 'Preliminary Report of the Special Rapporteur of the Commission on Human Rights on the Right to Food' (23 July 2001) UN Doc A/56/210, para 27 <www.un.org/unispal/document/auto-insert-187548/>.

²⁰ ibid, Ziegler.

for reasons beyond their control, to enjoy the right to adequate food by the means at their disposal. This obligation ensures that vulnerable groups such as victims of natural or other disasters or wars and crises continue to have access to food.²¹ States should realise the right to food progressively. The notion of ‘progressive realisation’ being a flexible device that recognises that full realisation of all rights in the ICESCR will generally not be achieved in a short period, cannot be interpreted as depriving the obligation of all meaningful content.²² Instead, it must be interpreted as imposing an obligation to move towards that goal as expeditiously and effectively as possible.²³

While the ICESCR recognises that state obligations to realise the right to food may be constrained by reasons beyond their control, such as lack of access to resources, the obligations do not abate even when states are faced with severe resource constraints such as climate change or economic crises or war, particularly concerning vulnerable individuals and groups.²⁴ The ICESCR prohibits the use of food as a tool for political pressure, and it is a violation of the right to discriminate in access to food on the grounds of race, colour, sex, language, religion, political or other opinions, national or social origin, property, birth or another status.²⁵ Compliance with the ICESCR is primarily by state parties, and persons or groups who are victims of the violation of the right are entitled to remedies under national and international laws.²⁶ States generally have an obligation to self-monitor and initiate strategies, including corrective legislation, to realise the right.²⁷ However, since 2000, the UN Special Rapporteur on the Right to Food has monitored states’ compliance with obligations under the ICESCR and SDG2 to eradicate hunger.

As noted above, this seeming robust legal framework has not translated into significant success for the right to food. One challenge is the susceptibility of food production to natural and environmental conditions.²⁸ Climate change, for instance, reduces crop yield, alters temperature, rainfall patterns, and already causes a loss of about 35 trillion consumable food calories per year.²⁹ Other contributory factors

²¹ Comment No 12, para 15.

²² *ibid* para 14; see also ICESCR, art 2(1); cf art 2 of the International Covenant on Civil and Political Rights (ICCPR) which embodies an immediate obligation to respect and ensure all the relevant rights.

²³ See ICESCR, art 11(2); Comment No 12, para 14; CESCR Comment No 3 ‘The Nature of States Parties’ Obligations (Art 2, Para 1 of the Covenant)’ (14 December 1990) UN Doc E/1991/23 (Comment No 3), para 9.

²⁴ *ibid*, Comment No 3 para 10; see also Comment No 12, paras 17, 28.

²⁵ ICESCR, art 2(1).

²⁶ Comment No 12, paras 32–35; see also *Social and Economic Rights Action Center and Center for Economic and Social Rights v Nigeria* Communication No 155/96 (African Commission on Human and People’s Rights, 27 May 2002); *People’s Union for Civil Liberties v Union of India* (1997) 1 SCC 301.

²⁷ *ibid*, Comment No 12, para 31.

²⁸ See FAO, ‘The Impact of Disasters and Crises on Agriculture and Food Security (2021)’ <<https://doi.org/10.4060/cb3673en>>.

²⁹ <<https://nca2018.globalchange.gov/chapter/10/>>. P Gowda and others, ‘Agriculture and Rural Communities’ in D R Reidmiller and others, (eds) *Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, Volume II* [2018] US Global Change Research Program, Washington, DC, USA 391.

are increasing urbanisation and population, which put pressure on available land and other resources for agriculture,³⁰ and changing food preferences, which require food production and FSCs to be more transparent.³¹ Section 3 shows how the value propositions of smart farms map to the core content of the right to food and the obligations of states. In particular, it improves food production methods, conservation, and distribution which is, in itself, an essential component for realising the right to food³² and overcoming the challenges discussed in this chapter.

3 Smart Farming

Smart farming entails collecting data, analysing the data, and making predictions for optimal management of crops and livestock. Smart devices connected to the internet control the farm using AI capable of executing autonomous actions or doing so remotely.³³ Big Data, sensors, IoT devices, agricultural robots (agribots), autonomous vehicles (AV) (eg tractors), cloud infrastructure, and AI play critical roles in smart farm systems. AI improves processes by leveraging multi-source data, learning, by detecting patterns in data, and making predictions to optimise outcomes for food production. In other words, AI can aggregate agronomic and weather data from sensors and IoT to make decisions about input such as water, fertiliser, and pesticide and predictions about planting, harvesting, and crop yield. It can provide real-time insights for operational decision-making and help to deliver ultra-precise optimisation of input resources.³⁴ Unlike precision agriculture, which often only takes in-field variability into account, smart farming bases decision-making on both location and data-enhanced context and situation awareness triggered by real-time events.³⁵ Examples of use cases include machine vision systems to recognise crop diseases, pest damage, and livestock diseases. Algorithms are trained to analyse images from cameras and drones, recognise specific crop, fruit, or animal condition patterns, and launch alerts or recommend remedial actions. Robots can identify and pick fruits without damaging them using sensors and cameras that enable decision-making in real-time.³⁶ Self-driving tractors and drones perform automated grassland mowing and planting, weed removal, and

³⁰ The UN estimates that, by 2050, 68 per cent of the world will live in densely populated urban areas—up from 55 per cent currently—and that global food production will need to increase by 70 per cent to meet a population estimated to increase to 9 billion people by 2050. See FAO (n 28).

³¹ See FAO (n 28) 70–82.

³² ICESCR, art 11(2)(a).

³³ Sjaak Wolfert, Cor Verdouw, and MJ Bogaardt, ‘Big Data in Smart Farming’ (2017) 153 Agricultural Systems 69–80, 70.

³⁴ *ibid.* See also the chapter by Alberto Quintavalla in this volume.

³⁵ *ibid.*

³⁶ See eg Nick Lavars, ‘Apple Harvesting Robot Plucks a Piece of Fruit Every 7 Seconds’ (*Newatlas*, 27 April 2021) <<https://newatlas.com/robotics/apple-harvesting-robot-fresh-seven-seconds/>>.

targeted pesticide application using autonomous navigation systems and sensors.³⁷ Predictive analytics can leverage ‘big farm data’³⁸ to generate insights on crop and livestock diseases, make price predictions, and provide market guidance.³⁹ When fully deployed, cloud-based platforms can connect, monitor, and automate the activities of a fleet of robots and feed data streams back to central systems.⁴⁰

Greenhouses and vertical and hydroponic farming are particularly significant use cases. For example, AI analyses agronomic and related data collected by IoT for decision-making on parameters like light, humidity, nutrients, and temperature thereby making it possible for autonomous greenhouses to deliver irrigation, scout for pests, and deliver precise pesticides.⁴¹ Operating at scale, greenhouses and hydroponic farms can wean farming off its dependency on land and weather, thus helping countries with little arable land and high dependency on food imports to accelerate food production.⁴²

In addition to the above, smart farms can increase FSC efficiency. FSCs consist of numerous participants and intermediaries—including farmers, processors, distributors, retailers, and consumers—and involve complex logistics of getting food to the right place at the right time and in the right quantity and quality. FSCs have substantial inefficiencies, impacting all actors in the chain, from producers to consumers. They are characterised by timeline delays that cause loss of nutrition and waste and are primarily opaque, leading to food fraud, contamination, health risks, and waste.⁴³ AI can predict yield shortages by evaluating data from multiple sources on smart farms, thus reducing emergency sourcing and potentially hunger. It can reduce risks to timelines, food quality, and health by alerting suppliers and buyers about quality breaches and identifying and predicting the path of food-borne diseases.⁴⁴ AI can also reduce food waste. Already, one-third of global food production (about 1.3 trillion tonnes worth around \$1 trillion) is wasted.⁴⁵ Finally,

³⁷ P Londjani and others, ‘AIA: AI and EU Agriculture’ (EC Science and Knowledge Service Joint Research Centre 2020) 12; see also Subhajit Basu and others, ‘Legal Framework for Autonomous Agricultural Robots’ (2020) 35 *AI & Society* 113.

³⁸ This is not a term of the art but could be used to describe Big Data in agriculture context. It would include agronomic, weather, and machine data and data on farming methods and practices particularly from different farms uploaded to a central system for analytics.

³⁹ Kiran M Sabu and TKM Kumar, ‘Predictive Analytics in Agriculture: Forecasting Prices of Areca Nuts in Kerala’ (2020) 171 *Procedia Computer Science* 699.

⁴⁰ Rich Hardiy, ‘Remotely Operated Spot Robot Herds Sheep in New Zealand’ (*New Atlas*, May 2019) <https://newatlas.com/robotics/robot-dog-spot-boston-dynamics-rocos-new-zealand-farm/?itm_source=newatlas&itm_medium=article-body>.

⁴¹ NOKIA, ‘Real Action: Smart Agriculture’ <www.nokia.com/networks/real-action/smart-agriculture/>.

⁴² Eg the United Arab Emirates (UAE) and Singapore import 90 per cent of their food because of land constraints.

⁴³ FAO, ‘2021 The State of Food and Agriculture: Making Food Systems More Resilient to Shocks and Stresses’ (2021) 48–61 <www.fao.org/3/cb4476en/cb4476en.pdf>.

⁴⁴ See eg Ilianna Kollia, Jack Stevenson, and Stefanos Kollias, ‘AI-enabled Efficient and Safe Food Supply Chain’ (MDPI, 2021) <<https://arxiv.org/pdf/2105.00333.pdf>>.

⁴⁵ UN SDG12: ‘Ensure Sustainable Consumption and Production Patterns’ <www.un.org/sustainabledevelopment/sustainable-consumption-production/>.

through end-to-end (or ‘farm-to-fork’) visibility, AI can help consumers evaluate label claims about provenance or the origin of food and how it was produced and transported.⁴⁶ Thus, it can empower consumers to actualise the satisfaction of dietary needs and fitness of food for sociocultural contexts as positive qualitative aspects of the right to food.⁴⁷ Better outcomes are expected as AI converges with distributed ledger technology (DLT) to increase food traceability and transparency and reduce food fraud, information asymmetry, and waste.⁴⁸

Smart farms can ultimately facilitate the expeditious progressive realisation of the right to food and the obligation of states to fulfil the right. However, AI in agriculture also creates some challenges that states must correspondingly address.⁴⁹ In sections 3.1 and 3.2, it is argued that AI in agriculture can sustain food and other systemic inequalities and create conflicts between the right to food and rights to privacy and decent work.

3.1 (Food) Inequality

Analytics and predictive models in smart farm systems need large volumes of data and extensive IoT infrastructure to train the algorithms. However, there are severe gaps in available agronomic data needed to develop AI systems. In low- and middle-income countries (LMIC), smallholder farmers, women, and indigenous farming communities who play critical roles in food production, tend to have little or no historical data, and a sparse digital footprint. This can severely limit the development and functionality of AI systems in LMICs. Also, while Big Data analytics, cloud-based storage, satellite imagery, remote sensors, and mobile connectivity have made AI increasingly feasible in agriculture, smallholder farmers still face high capital inputs and skills deficits.⁵⁰ The same group, particularly in LMICs, also lacks access to credit and other value added financial services such as insurance. As big farm data is increasingly used for alternative credit scoring, such farmers may be further marginalised, making it impossible to adopt or scale the technology. AI may skew food production at the global level, favouring advanced economies with access to data, infrastructure, research, and funds. For example, while more than one-third of the planet’s undernourished (282 million) live

⁴⁶ Kollia, Stevenson, and Kollias (n 44).

⁴⁷ See ICESCR, art 11.

⁴⁸ FAO, ‘Opportunities for the Application of Blockchain in the Agri-Food Industry’ (revised edn July 2020) <www.fao.org/3/ca9934en/CA9934EN.pdf>.

⁴⁹ Sandy De Alwis and others, ‘A Survey on Smart Farm Data, Application and Techniques’ (2022) 138 Computers in Industry 103624.

⁵⁰ International Finance Corporation (IFC), ‘Artificial Intelligence in Emerging Markets: Opportunities, Trends and Emerging Business Models’ (2021) 72 <www.ifc.org/wps/wcm/connect/95a40480-27b5-4b99-8b4c-7768ae6a53a2/AI-Report_Web.pdf?MOD=AJPERES&CVID=nhLrsRc>.

in Africa, the region is also significantly underdeveloped in application, research, and innovation of agritech.⁵¹ In contrast, Europe and the United States (US) are the most viable markets for investment in smart farms. The problem here is paradoxical and complex; some parts of the world may not produce enough food, for example, because they lack access to arable land, or state of the art technologies, while others may over-produce. Section 3 already highlighted the inefficiency of FSCs and their propensity for waste. This may be worsened by over-production.

Nevertheless, in enhancing FSCs efficiency, data from smart farms will be shared among multiple stakeholders. Some data, such as those relating to farming and processing methods, or livestock rearing, may be proprietary. There is a risk that farmers will lose control over such proprietary information⁵² or be exploited or disadvantaged by data sharing. For instance, in *Haff Poultry Inc et al v Tyson Foods Inc*,⁵³ broiler chicken growers brought a class action against the defendants for allegedly sharing their production data, including grower payments, broiler weights, types of feed and medicine used, and transportation costs, with third parties. The defendants belong to a cartel who shared the data intending to keep growers' compensation below the competitive levels. The action was adjudged anti-competitive, predatory, and unfair.

While the case does not involve AI, it underscores the concerns about 'AI colonialism' in the broader debate on the impacts of AI on societies. AI colonialism is a neocolonial concept characterised by taking power and resources in the form of data from marginalised communities to profit the already wealthy.⁵⁴ Data available and collected from one country or community is labelled and used to train algorithms in another, the AI systems are developed in yet another country, while the economic benefits of the ecosystem accrue to another country.⁵⁵ Applied to the smart farm context, those who create big farm data, like farmers and their communities, may not directly benefit from it, either in the form of profit or infrastructure. Instead, the main beneficiaries of the data will be global corporations, (middlemen) data companies, and their countries of registration and taxation. Generally, therefore, AI may widen digital and wealth gaps, foster food dependency, and increase states' vulnerability. For food security and sustainability, the resulting trust deficit here can lead to data silos that impede the flow of data needed to improve interoperability and build transformative agricultural business models

⁵¹ Eg, FAO, 'The Future of Food and Agriculture: Trends and Challenges (2017)' 123–30 <<https://www.fao.org/3/i6583e/i6583e.pdf>>.

⁵² Mannak Gupta and others, 'Security and Privacy in Smart Farming: Challenges and Opportunities' (2020) 8 IEEE Access 34569.

⁵³ *Haff Poultry Inc et al v Tyson Foods Inc*, Case No 6:17-CV-00033-RJS (US District Court for the Eastern District of Oklahoma, 2017).

⁵⁴ Stanford University Human Centred AI (HAI), 'The Movement to Decolonise AI: Centering Dignity over Dependency' (21 March 2022) <<https://hai.stanford.edu/news/movement-decolonize-ai-centering-dignity-over-dependency>>.

⁵⁵ Karen Hao, 'AI is Creating a New Colonial World Order' (*MIT Tech Review*, 19 April 2022) <<https://www.technologyreview.com/2022/04/19/1049592/artificial-intelligence-colonialism/>>.

for efficient food production and distribution. As one recent survey of Australian farmers suggests, farmers may refuse to share data if they do not trust the system.⁵⁶

3.2 Conflicting Rights

Personal data is perhaps the most strictly regulated among the diverse data collected by smart farms because of its association with privacy.⁵⁷ Personal data is defined as information relating to an identifiable natural person and typically includes names, addresses, geolocation, and any data that makes identification possible.⁵⁸ Data processing is governed by principles of lawful processing, purpose specification and data minimisation, and accuracy, security, and transfer.⁵⁹ As indicated above, sharing big farm data can benefit the entire food value chain, while correspondingly disadvantageous to farmers. In this case, it can violate farmers' privacy and data protection rights of other stakeholders. To illustrate, software licences embedded in farm equipment, while notoriously lengthy and difficult to understand, can trigger a broad range of data collection and sharing.⁶⁰ Smallholder farmers may operate as sole traders rather than companies, making it easier to link farm data directly to individual farm owners, their workers, visitors, suppliers, and customers. Even when data is supposedly anonymised, AI algorithms can track data back to its source farm and invariably lead to the identification of individual farmers. Inferences about income and financial status can be drawn from product yield predictions and data collected by machines and sensors, with potentially discriminatory outcomes.⁶¹

Notably, states obligation under the ICESCR extends to the protection of farmers' privacy and their protection from exploitation. States parties should pursue policies that eliminate public or private discriminatory practices in the realisation of civil and political rights, including the right to food, and ensure that activities of private actors conform with the right to food and other human rights.⁶²

It is important to mention that privacy and data protection requirements can also have adverse impacts on the realisation of the right to food. Indeed, protecting personal data in the context of the right to food implies compliance with legal and

⁵⁶ See Leanne Wiseman and others, 'Farmers and their Data: An Examination of Farmers' Reluctance to Share their Data through the Lens of the Law Impacting Smart Farming' (2019) 90–91 NJAS Wageningen Journal of Life 100301.

⁵⁷ See eg UDHR, art 12; EU Charter of Fundamental Rights, arts 7, 8.

⁵⁸ See eg GDPR (2016), art 4(1); GDPR Convention 108+ (2018), art 2(a).

⁵⁹ See eg GDPR, art 5. See also the chapter by Alessia Zornetta and Ignacio Cofone in this volume.

⁶⁰ Wiseman and others (n 56).

⁶¹ Eg by perpetuating existing inequality in access to credit by smallholder farmers, see eg Aaron Klein, 'Reducing Bias in AI-Based Financial Services' (*Brooking Report*, July 2020) <www.brookings.edu/research/reducing-bias-in-ai-based-financial-services/>.

⁶² Comment No 12, para 27.

bureaucratic processes that can create delays in identifying beneficiaries of food aid during humanitarian crises.⁶³

Finally, smart farms can create conflicts between the rights to food and work. The right to work covers just and favourable conditions of work and a right to fair wages, and states are obliged to ensure the existence of services to help people identify employment opportunities and find work.⁶⁴ Conversely, innovation tends to create redundancies, drive down wages for low skilled workers and create opportunities for highly skilled workers. Thus, in poor economies or countries with high dependency on agriculture, smart farms can worsen unemployment problems. Rural labour declines and unemployment may be higher than urban job opportunities created by technology.⁶⁵ Smart farms can also undermine efforts to reduce poverty, hunger, and undernourishment as job losses and lower wages impair economic access to food that is nutritional and fit for a people's sociocultural context. Overall, while job losses to automation in agriculture may not be entirely negative because agriculture has a higher dependency on repetitive and tedious manual labour and traditionally suffers labour shortages, it can exacerbate unemployment, social mobility, and income disparities.

4 Conclusion

As the environmental, climate, and other factors like war, pandemics, and civil unrest increase the threats to food security, a data-driven approach can make the difference between hunger and undernourishment and food security and sustainability. From the right perspective, the legal framework is robust. The law already anticipates using technology and urges its application to realise the right to food. However, challenges are likely to undercut the positive effects of AI application in agriculture. States are implicated in addressing these challenges. An inclusive approach to the development and deployment of AI is needed. States have an obligation under the ICESRC to fulfil the right to food by facilitating access to resources for food production, in this case, technology, particularly for smallholder farmers, women, indigenous communities, and other vulnerable stakeholders. State parties also have extraterritorial obligations—to pursue international cooperation and assistance in the realisation of the right to food—which can translate into advanced economies providing affordable and accessible AI infrastructure and promoting equity in the sharing of benefits accruing from broader use of big farm data. Poorer

⁶³ This point is not discussed fully in this chapter, however, for better understanding of the World Food Programme (WFP) 'Guide to Privacy and Data Protection' (June 2016) <<https://docs.wfp.org/api/documents/e8d24e70cc11448383495caca154cb97/download/>>.

⁶⁴ ICESCR, art 6.

⁶⁵ See eg Luc Christiaensen and others, 'Viewpoint: The Future of Work in Agri-food' (2021) 99 Food Policy 1.

states can develop national strategies to prioritise investment in research and real-world applications of AI for sustainable food production. All states must adopt appropriate measures, including policy and legislative frameworks that protect human rights principles in an AI context. Policies must identify potential risks in using AI in agriculture and develop responses to address them.

Rights are rarely absolute, and there is no suggestion that the right to food trumps—or should be achieved at the expense of—other rights. In fact, as stated in Comment No 12, national strategies to protect the right to food must include measures to respect and protect self-employment and decent living for wage earners and their families.⁶⁶ Conversely, the right to food is justiciable, and courts have intervened to enforce the right and compensate the victims of states' failure to respect, protect, and fulfil the right to food. Therefore, states must do some strict balancing to avoid or limit conflicts between different rights in the realisation of the right to food. Laws may promote responsible and right-respecting innovations by categorising discriminatory AI systems as potentially harmful. Templates can be developed for more transparent contracts articulating data ownership, liability for misuse and exploitation of big farm data, and sharing of benefits from the value generated by data. This approach creates certainty, fosters trust, and liberalises data sharing for food security. For instance, privacy policies are notoriously complex and farmers can better understand contractual terms and conditions under which personal and farm data are used when agreements are standardised. Finally, measures to protect workers and ensure decent wages may include re-training individuals and groups who are likely to be made redundant by the introduction of the technology discussed in this chapter.

⁶⁶ ICESCR, art 11; Comment No 12, para 12.

24

Artificial Intelligence and the Right to Housing

Caroline Compton and Jessie Hohmann

1 Introduction

The aims and implications of artificial intelligence (AI)¹ and its applications to housing are global in scope,² as are the aims and implications of human rights, which are embedded into international law as global standards. Moreover, the applications of AI to housing are potentially vast, spanning from the panoramic to the granular. The intersections between AI and housing cover areas as wide as real estate finance and the (broader) finance and insurance industry; aspects of the built environment, from urban planning to refuse collection to transport; to management and intervention at the level of the individual dweller (with a myriad of tiny daily impacts in, over and around the home) such as control of thermostats and smart doorbells. There is no aspect of housing that AI might not impact, given the scope of its applications and rapidity of its development. For these reasons, an evaluation of the intersections between AI and the right to housing is both pressing and timely.

Although national and supranational law-making bodies are actively considering the implications of AI for human rights, the impacts on the right to housing have not, so far, been addressed in any detail.³ Scholars are also only beginning to discuss AI and the right to housing.⁴ This is also an area that is, so far, underexplored

¹ For the purpose of this chapter, we treat AI as being automated decision-making (ADM) or alerts based on a range of different computer methodologies. See also the chapter by Martina Šmuclerová, Luboš Král, and Jan Drchal in this volume.

² See Mara Ferreri and Romola Sanyal, 'Digital Informalisation: Rental Housing, Platforms, and the Management of Risk' (2022) 37(6) *Housing Studies* 1035.

³ See eg European Union Agency of Fundamental Rights, 'Getting the Future Right: Artificial Intelligence and Fundamental Rights' (European Agency of Fundamental Rights 2020); European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

⁴ See eg Yin Xu and Hong Ma, 'Research and Implementation of the Text Matching Algorithm in the Field of Housing Law and Policy Based on Deep Learning' (2021) Complexity 1–9; Joshua Chad Gellers and David Gunzel, 'Artificial Intelligence and International Human Rights Law: Implications for Humans and Technology in the 21st Century and Beyond' in A Zwitter and O Gstrein (eds), *Handbook on the Politics and Governance of Big Data and Artificial Intelligence* (Edward Elgar 2023). See

in United Nations (UN) human rights standard setting and other benchmarking or monitoring processes. For example, the UN Special Rapporteur on Adequate Housing's recent report on housing discrimination and spatial segregation does not discuss the impact of AI,⁵ despite AI's deep implication in housing discrimination as discussed below, although the UN Special Rapporteur on Extreme Poverty and Human Rights is attuned to the issue in the broader social welfare context.⁶

While much of the literature on housing and AI focuses on ways that AI can negatively impact experiences of housing and home, in this chapter we also wish to draw readers' attention to AI as means through which enjoyment of the right to housing can be advanced. We begin the chapter by defining the right to housing—concentrating specifically on the right under the International Covenant on Economic, Social and Cultural Rights (ICESCR), as the international standard for the right, and—albeit to a lesser extent—the right under the Revised European Social Charter given the vigorous debates over balancing AI and human rights in Europe. We then turn to discuss first, uses of AI that impact negatively on the right to housing; and second beneficial applications of AI for the right to housing.

Our conclusion is that AI technologies have both positive and highly problematic applications in the field of housing. While many of the technologies may themselves be neutral,⁷ their use will unfold in a world of pre-existing unequal power relations, biases, and structural inequalities. As such, their uses may very often exacerbate these inequalities and biases, leading to the violation of human rights, including the right to housing. However, human rights law (HRL) provides standards through which we can attempt to regulate such technologies and provide a framework of principles against which to evaluate them.

2 Defining the Right to Housing in International Law

The right to housing has been recognised as part of international human rights law (IHRL) since its inclusion in the 1948 Universal Declaration of Human Rights (UDHR). It is now codified in, or implied into, several international and regional

also 'Project Evict', which is using data analysis to gather information on evictions and connect it to an analysis of the right to housing <www.eviction.eu/data-science-eviction-law/>.

⁵ See Report of the Special Rapporteur on adequate housing as a component of the right to an adequate standard of living, and on the right to non-discrimination in this context: Balakrishnan Rajagopal, 'Discrimination in the Context of Housing' (14 October 2021) UN Doc A/76/408.

⁶ See Brief by UN Special Rapporteur on extreme poverty and human rights as Amicus Curiae in the case of *NJCM cs v De Staat der Nederlanden* (SyRI) before the District Court of the Hague C/09/550982/HA ZA/18/388.

⁷ But see the longstanding debate about the politics of technologies, eg Langdon Winner, *The Whale and the Reactor: A Search for Limits in an Age of High Technology* (University of Chicago Press 1986); Merrit Roe Smith and Leo Marx (eds), *Does Technology Drive History?: The Dilemma of Technological Determinism* (MIT Press 1994).

human rights treaties,⁸ and numerous national constitutions.⁹ The international standard is the ICESCR, ratified by over 170 states. This treaty includes the right to adequate housing—more than mere shelter—as part of the broader right to an adequate standard of living, in article 11(1):

The States Parties to the present Covenant recognize the right of everyone to an adequate standard of living for himself and his family, including adequate food, clothing and *housing*, and to the continuous improvement of living conditions.

The Committee on Economic, Social and Cultural Rights (CESCR or Committee) which oversees the implementation of the ICESCR by state parties, has defined the right as, at root, ‘a place to live in security, peace and dignity’.¹⁰ This means that the right is about more than four walls and a roof. Adequate housing connects to community, to rights to political and public life, to self-expression (even self-determination), and privacy.¹¹ In its authoritative statement on the content and scope of the right, the Committee has stated that adequate housing is made up of seven essential elements, all of which must be present to at least the level of a minimum core. These are: legal security of tenure; availability of services, materials, facilities, and infrastructure; affordability; habitability; accessibility; location; and cultural adequacy.¹² Implicit within article 11(1)—and explicit within ICESCR as a whole¹³—is the requirement for non-discrimination in housing.

States’ obligations for the right (in article 2(1) of the ICESCR) further clarify its scope. These obligations include a mix of immediate and longer-term obligations arising from article 11 and article 2(1). Immediate obligations include those which do not entail major resources, such as the repeal of discriminatory laws, and the regulation of private sector actors, such as the real estate and construction

⁸ These human rights instruments are discussed in more detail in Jessie Hohmann, *The Right to Housing: Law, Concepts, Possibilities* (Hart 2013) chs 1–3.

⁹ See M Oren and R Alterman, ‘The Right to Adequate Housing Around the Globe: Analysis and Evaluation of National Constitutions’ in S Agarwal (ed), *Rights and the City: Problems, Progress, and Practice* (University of Alberta Press 2022). The most well-known and influential of these are the South African and Indian Constitutions, discussed further in Hohmann (n 8) ch 4.

¹⁰ UN CESCR, ‘The Right to Adequate Housing (Article 11(1)): General Comment 4’ (1991) UN Doc E/1992/23 (General Comment No 4), para 7.

¹¹ See further Hohmann (n 8) Pt II (on privacy, space, and identity). On self-determination, see ‘Report of the Special Rapporteur on Adequate Housing as Component of the Right to an Adequate Standard of Living, and on the Right to Non-Discrimination in this Context’ (17 July 2019) UN Doc A/74/183.

¹² General Comment No 4, para 8; see for further discussion Hohmann (n 8) 20–29.

¹³ ICESCR, art 2(2). Non-discrimination is also a norm of customary international law, and as such not subject to the standards of progressive realisation contained in the ICESCR, but rather imposes an obligation of immediate fulfilment. On immediate obligations and discrimination see UN CESCR ‘General Comment No 20: Non-discrimination in Economic, Social and Cultural Rights (art 2, para 2 of the International Covenant on Economic, Social and Cultural Rights)’ (2 July 2009) UN Doc E/C.12/GC/20, para 8.

industries,¹⁴ and enabling people to access adequate housing without undue legal and political barriers. Beyond immediate obligations, the ICESCR requires states must move toward *full* realisation of the right for all, using the maximum available resources at their disposal. To do so may require positive action such as subsidising housing for those unable to access it in the market, and more broadly ensuring a social system in which the vulnerable and marginalised are able to live in peace, dignity, and security. This in turn may require broader social welfare measures (such as social security) the protection of other social rights (such as the right to decent work) and steps toward a more just and equitable society.¹⁵

In Europe,¹⁶ the right to housing is also recognised under the Revised European Social Charter (Social Charter) in article 31, which the European Social Committee has interpreted robustly. This right imposes three obligations. First, under article 31(1), the state is to promote access to housing that is of an acceptable standard. The second, under article 31(2), is an obligation for the prevention of homelessness, and its reduction over time, with the ultimate aim being its elimination. The third obligation, corresponding to article 31(3), is specifically concerned with affordability for those without adequate resources.

The Social Committee has taken notable steps to explain the scope of the right to housing in its collective complaints jurisprudence.¹⁷ It has defined adequate housing as '[a] dwelling which is safe from a sanitary and health point of view, that is, possesses all basic amenities, such as water, heating, waste disposal, sanitation facilities, and electricity; is structurally secure; not overcrowded; and with secure tenure supported by law'.¹⁸ In its decisions, the Committee has also defined affordable housing¹⁹ and has noted that affordability should not be measured with

¹⁴ On obligations to regulate private actors and the repeal of discriminatory laws, see 'Report of UN Special Rapporteur on Adequate Housing as a Component of the Right to an Adequate Standard of Living, and on the Right to Non-Discrimination in this Context' (14 October 2021) UN Doc A/76/408 para 9. For a detailed treatment of obligations under the ICESCR, see M Sepúlveda Carmona, *The Nature of the Obligations Under the International Covenant on Economic, Social and Cultural Rights* (Intersentia 2003).

¹⁵ Hohmann (n 8) ch 1.

¹⁶ Note also the right to housing assistance in the EU Charter of Fundamental Rights. See Charter of Fundamental Rights of the European Union [2012] OJ C326/02, art 34(3).

¹⁷ The Committee has implied the standards of adequate housing into other provisions of the Social Charter that have a more marginal reference to housing, including art 16's protection of the family. See *European Roma Rights Centre (ERRC) v Greece*, Complaint no 15/2003, decision on the merits of 7 February 2005, para 17; Centre on Housing Rights and Evictions (COHRE) v Italy, Complaint no 58/2009, decision on the merits of 25 June 2010, para 115. This bold interpretive move has made an 'almost peripheral' reference to housing a central housing rights provision under the Social Charter. See Urfan Khaliq and Robin Churchill, 'The European Committee of Social Rights: Putting Flesh on the Bare Bones of the European Social Charter' in Malcolm Langford (ed), *Social Rights Jurisprudence: Emerging Trends in International and Comparative Law* (CUP 2008) 429, 448.

¹⁸ *European Federation of National Organisations Working with the Homeless (FEANTSA) v France*, Complaint no 39/2009, decision on the merits of 5 December 2007, para 76. See also *ERRC v Greece* (n 17) para 16. This has been expanded to include access to fresh water. *European Roma Rights Centre v Portugal*, Complaint no 61/2010, decision on the merits of 30 June 2011, para 36.

¹⁹ *FEANTSA v France* (n 18) para 124; *COHRE v Italy* (n 17) paras 41–42.

reference to the average person, but the poorest.²⁰ With regard to eviction and the right to housing, the Committee has held that evictions must be undertaken ‘in conformity with the dignity of the persons concerned’.²¹ Procedural guarantees are necessary.²²

The Committee has clarified that the state must take a number of steps to demonstrate compliance with the right under the Social Charter.²³ Like obligations under the ICESCR, states parties are required to move toward full realisation of adequate, affordable housing, and ensure that their legislative and policy frameworks promote access to housing for all. Where people are unable to access housing in the market, the state may need to subsidise or otherwise provide housing through positive measures.

With a clearer picture of the right to housing’s scope and the obligations it entails, we now turn to discuss AI and its applications in the sphere of housing, specifically discussing their implications for the right to housing.

3 The Application of AI in the Housing Sphere: Emerging Violations

It is becoming increasingly clear that AI and its uses in the field of housing can have discriminatory effects, which undermine efforts to realise a right to housing. Among these are platforms which connect landlords and tenants, or housemate with housemate; services that offer automated tenant screening services, even automated evictions; and practices of ‘algorithmic redlining’. These specific examples, which we discuss further below, sit within a broader global financial sector—a political economy—which also increasingly relies on AI. For example, a range of platforms seek to aid those with capital wishing to invest it in real estate. These are normally geared to large investors or funds but can have smaller scale applications too.²⁴ These AI tools are engaged at multiple scales, in service of the financialisation of real estate.²⁵ Financialisation itself, as the UN Special Rapporteur on Housing has pointed out, is detrimental to the ability of ordinary people—especially those with low income—to access adequate housing.²⁶ The linking of housing into

²⁰ European Federation of National Organisations Working with the Homeless (FEANTSA) v Slovenia, Compliant no 53/2008, decision on the merits of 8 September 2009, para 72.

²¹ COHRE v Italy (n 17) para 67.

²² Centre on Housing Rights and Evictions (COHRE) v France, Complaint no 63/2010, decision on the merits of 28 June 2011, para 41–42.

²³ FEANTSA v France (n 18) para 56.

²⁴ Joe Shaw, ‘Platform Real Estate: Theory and Practice of New Urban Real Estate Markets’ (2020) 41 *Urban Geography* 1037, 1048–49.

²⁵ Desiree Fields and Dallas Rogers, ‘Towards a Critical Housing Studies Research Agenda on Platform Real Estate’ (2021) 38 *Housing, Theory and Society* 72, 81.

²⁶ ‘Report of the Special Rapporteur on Adequate Housing as Component of the Right to an Adequate Standard of Living, and on the Right to Non-Discrimination in this Context’ (18 January 2017) UN Doc A/HRC/34/51.

broader global financial circuits, for the purpose of making money for investors in those circuits, is thus a challenge for the right to housing generally.²⁷ One of the significant roles played by AI here is in accelerating the rate of transactions,²⁸ through which AI further attenuates the relationship between housing and its use value, as a place to live in peace, dignity, and security. The human-centred basis of human rights insists instead on the importance of the relationship between the person and the home as living space.

'Algorithmic redlining' is also of serious concern. It both sits within and perpetuates a financialised housing system. This practice occurs where Big Data use by both public and private actors perpetuates the practice of excising minorities from access to housing credit or finance.²⁹ Algorithmic redlining relies on pre-existing discrimination—such as the data from historic redlining. At the same time, its reliance on Big Data and automated decision-making (ADM) from other areas compounds the discrimination, as it gathers in other discriminatory scoring and targeting technologies,³⁰ such as those discussed further below. Algorithmic redlining removes accountability on the basis of discriminatory intention, instead dressing discriminatory effects in the legitimating clothes of objectivity, even if the criteria for decisions are unknown, as they are when neural nets are used. A 2020 United States (US) case, *Connecticut Fair Housing Center v CoreLogic Rental Property Solutions*, serves as an example. The applicant challenged the decision, made through an automated screening process, to deny her disabled son access to housing because of a shoplifting charge brought against him—and dropped—before his disability. The Court was unable to find that CoreLogic's algorithmic decision-making tools either did or did not breach the Fair Housing Act, even while finding that the application used, CrimSAFE, 'may be, but is not necessarily as a matter of law, a proximate cause of housing discrimination'.³¹ In other words, the Court chose not to look inside the 'black box' of the AI to determine if discrimination did in fact occur, even when the effect appeared discriminatory.³²

Another way in which AI can work against realising the right to housing is in the realm of a number of services that use Big Data, sourced from social media, rental contracts, and the supporting documentation required to enter a tenancy (such as bank statements), on a host of platforms that offer to match tenants with landlords

²⁷ ibid.

²⁸ Michael A Peters and Tina Besley, 'Critical Philosophy of the Postdigital' (2019) 1 Postdigital Science and Education 29; Sarah Keenan, 'From Historical Chains to Derivative Futures: Title Registries as Time Machines' (2019) 20 Social & Cultural Geography 283.

²⁹ James A Allen, 'The Color of Algorithms: An Analysis and Proposed Research Agenda for Deterring Algorithmic Redlining' (2019) XLVI Fordham Urban Law Journal 219.

³⁰ ibid discussing the interaction between housing finance and, for example, payday loans.

³¹ *Connecticut Fair Housing Center v CoreLogic Rental Property Solutions* [2020] District Court of Connecticut 3:18-CV-705 (VLB).

³² For a discussion of the Court's difficulty in examining the 'black box' in the Dutch *SyRI* case, see A Rachovitsa and N Johann, 'The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch *SyRI* Case' (2022) 22(2) Human Rights Law Review 1–15, 11.

and housemate with housemate. A suite of highly profitable companies provides automated credit, reference, and employment checks derived from a range of data sources to landlords.³³ For those whose life trajectory reflects dominant narratives of ‘good’ or acceptable behaviour, the automation of these checks can be straightforward, albeit significantly increasing the cost of a rental transition. For example, one website allows tenants to purchase a product called RentCheck, where tenants can ‘request to be endorsed as an awesome renter by your previous property managers’.³⁴ This renders something previously required as a matter of due diligence by new landlords into a product that must be purchased by the tenant. For renters, particularly those whose life trajectory differs from the mainstream, the loss of an opportunity to self-represent and explain can cost them the opportunity for housing. This is directly contrary to the right to housing requirement of accessibility for disadvantaged and marginalised groups.³⁵ These tenant screening services can clearly enable private parties and institutional landlords to discriminate against tenants on grounds prohibited under HRL (such as race), while simultaneously reinforcing structural biases and spatial segregation on a more systemic level.

Automated decision-making extends into the realm of software promising to manage all aspects of a tenancy, not only finding a tenant but automating maintenance and the ongoing landlord-tenant relationship. Some firms even promise automated eviction triggered by algorithm.³⁶ Such automated eviction contravenes the strict procedural requirements around eviction, as well as undermining security of tenure, both recognised as crucial elements of adequate housing in IHRL.³⁷

Finally, a host of technologies involving the ‘smart home’ and the internet of things can be used to surveil and manage tenants in ways that violate the home’s nature as a private space,³⁸ and that might undermine elements of the right to adequate housing including security of tenure, and cultural adequacy, for example, even as they may make housing more accessible, as we discuss further below.

These forms of AI and their discriminatory applications have concerning implications for the realisation of the right to housing. They deserve attention and critique, as well as regulation, to ensure that they are not used in contravention of the right. However, while they are the most commented on and visible of negative impacts on the right to housing, it is important to point out that AI’s applications in the field of housing are not confined to concerns about discrimination. AI can

³³ See eg ‘Renter Resume—Free Rental and Tenancy Application Form Online’ <www.rent.com.au/resume>.

³⁴ *ibid.*

³⁵ General Comment No 4, para 8(e).

³⁶ For a discussion, see Erin McElroy, ‘Property as Technology: Temporal Entanglements of Race, Space, and Displacement’ (2020) 24 *City* 112.

³⁷ See UN CESCR, ‘General Comment No 7: The Right to Adequate Housing (art 11.1): Forced Evictions’ (20 May 1997) UN Doc E/1998/22; General Comment No 4, para 8(a).

³⁸ For a comprehensive discussion of AI and privacy rights, see the chapters by Alessia Zornetta and Ignacio Cofone; and Natalia Menéndez González in this volume. See also Hohmann (n 8).

be used to remake the urban³⁹—and rural—infrastructure of our states changing the shape of cities and societies as a consequence. Accordingly, we must not only pay attention to the specific, but also to the background landscape of finance, infrastructure, and political economy, as noted at the beginning of this section.

4 AI as a Means Through Which to Realise the Right to Housing

Despite this concerning picture, it would be wrong to suggest that all AI works against the realisation of the right to housing. Acknowledging the importance of technological advance in helping improve the human condition, we turn to discuss three instances where AI might be used to realise or ensure the right.

4.1 Disaster Risk Reduction

Natural disaster presents a perennial risk to many billions of people,⁴⁰ a risk significantly exacerbated by anthropogenic climate change.⁴¹ Floods, droughts, bush or wildfires, and major storm events threaten housing stock, reduce the utility of areas, thus impacting access to, and the quality of, housing. Protecting the right to housing in the context of natural disasters has long been a focus of the UN Special Rapporteur on housing and this focus will only become more urgent in the face of climate change, rapid urbanisation, and population growth.⁴² The city of Jakarta, for example, is particularly vulnerable to the impact of flooding and has been subject to regular, major flooding events since the 1990s.⁴³ The city sits on a delta, which is highly polluted and many of the canals are lined with informal settlements.⁴⁴ The city's flooding is expected to become significantly worse in coming decades,⁴⁵ directly impacting the ten and a half million people who called the city home in 2020.⁴⁶

³⁹ For an argument that platform real estate is a quintessentially urban phenomenon, see Shaw (n 24) 1055–56.

⁴⁰ Peijun Shi and others, *World Atlas of Natural Disaster Risk* (Springer 2015).

⁴¹ Maarten K van Aalst and others, 'The Impacts of Climate Change on the Risk of Natural Disasters' (2006) 30 *Disasters* 5.

⁴² See Raquel Rolnik, 'The Human Right to Adequate Housing' in Flavia Zorzi Giustiniani and others (eds), *Routledge Handbook of Human Rights and Disasters* (Routledge 2018) 181; Miloon Kothari, 'Report of the Special Rapporteur on Adequate Housing as a Component of the Right to an Adequate Standard of Living, and on the Right to Non-Discrimination in this Context' (13 February 2008) UN Doc A/HRC/7/16, paras 2, 81–86.

⁴³ Christopher Silver, *Urban Flood Risk Management: Looking at Jakarta* (Routledge 2022).

⁴⁴ *ibid.*

⁴⁵ Hiroshi Takagi and others, 'Projection of Coastal Floods in 2050 Jakarta' (2016) 17 *Urban Climate* 135.

⁴⁶ Badan Pusat Statistik, 'Sensus Penduduk 2020–2020 Census Indonesia' <<https://sensus.bps.go.id/main/index/sp2020>>.

AI is being used to respond to this risk, particularly using social media and sensor-informed platforms that alert citizens and government to flooding events as they occur.⁴⁷ Developed following significant flooding in 2013 and 2014, are projects such as Peta Jakarta and PetaBencana.⁴⁸ PetaBencana, which utilises CogniCity software, pulls data from several sources. Using the Twitter API,⁴⁹ the platform contacts users who have geotagged a tweet with 'flood' or 'banjir' via a message, asking them to confirm if there is a flood in their area. Users are asked to provide information on specific location, depth of water, to upload an image, and any other information. The platform also pulls government GIS (geographic information system) data, and government field officers can add information, while also pulling data from government APIs streaming rainfall data and data from gauges on pumps, waterways, and floodways. This information is processed to present a map outlining flood risk to both public and government. The platform can then 'develop predictive tools to ready [Jakarta] for future disasters and become more resilient in day-to-day operations'.⁵⁰ The data can be used to inform city policy that responds to flood risk to residents' housing. This directly responds to the need for housing to be habitable—that is, safe for the residents to live in.⁵¹ It also may, indirectly and depending on state policy responses, respond to elements of location and of access to materials, services, facilities, and infrastructure.⁵² This is if responses take into account protecting current housing (and its location) through improved flood response infrastructure. In these ways, it can further enjoyment of the right to adequate housing.

On the other hand, the data may be used to displace residents in areas prone to flooding, without adequate consultation or attention to the communal aspects of living that the right to housing also seeks to protect. They may be displaced due to an assessment of high risk, or their homes may be removed to make place for disaster mitigation infrastructure that primarily services others. This links in with the gaps in the data that the platform captures. PetaBencana, like all data-processing tools, demonstrates inherent limitations to what data is utilised, and its accessibility and utility. Somewhere between 1–2 per cent of all tweets, worldwide, are

⁴⁷ See the chapter by Alberto Quintavalla in this volume.

⁴⁸ Peta Jakarta—the product of a partnership between the BPBD (Jakarta regional disaster management agency), Jakarta Province, Twitter, and the SMART Infrastructure Facility at the University of Woolongong PetaBencana—is being developed by the Urban Risk Lab at the Massachusetts Institute of Technology (MIT), which is furthering development of the CogniCity software platform.

⁴⁹ An API is an 'Application Programming Interface'. It is the point of contact between two applications. In this instance, the Twitter API is the interface that allows users to engage with the tweets posted to Twitter.

⁵⁰ 'CogniCity' (*Intelligent City Software and Solutions*, 15 September 2016) <<https://icos.urenio.org/applications/cognicity/>>.

⁵¹ General Comment 4, para 8(d).

⁵² *ibid* para 8(f) and (b).

geotagged.⁵³ This means most of all possibly relevant tweets are not captured by the PetaBencana platform. Given that using location services costs battery life (and when using apps, data), and accessing Twitter in the mode anticipated by the platform, a smartphone, we would imagine that areas frequented by poorer Jakartans are less represented on the platform. This is of particular concern because government attention flows because of platform suggestions, with the National Emergency Management Agency using the platform to inform actions.⁵⁴ This could result in potential protection of the housing of those who are better off, rather than according to actual housing need, and pushes against the right to non-discrimination in housing in post-disaster settings.⁵⁵ Nonetheless, while gaps in data creation might create imperfect maps, the enthusiastic uptake of the PetaBencana platform by the Indonesian government speaks to its utility. Flooding events can be highly localised and occur very rapidly, outflanking the capacity of government officials to respond in time. Real-time information improves decision-making and leads to high-quality predictions, improving disaster response, with the real potential to contribute to the enjoyment of the right to safe, habitable housing.

4.2 Remaining in Place

We turn now to our second example of how AI-utilising technology can improve the experience of, and access to, housing. Australia,⁵⁶ Europe,⁵⁷ the US,⁵⁸ and China⁵⁹—amongst others—are experiencing rapidly ageing populations. This presents several challenges for ensuring the right to housing for older persons, as well as persons with disabilities, both of whom the UN CESCR recognises as entitled to special consideration to ensure their right to housing.⁶⁰ Here, we focus on those who desire to remain living in their existing houses for as long as possible, despite requiring significant support to do so. We highlight how ambient AI-informed technology, that is, tech embedded into the house and operating

⁵³ Stephan Schlosser, Daniele Toninelli, and Michela Cameletti, ‘Comparing Methods to Collect and Geolocate Tweets in Great Britain’ (2021) 7 *Journal of Open Innovation: Technology, Market, and Complexity* 44.

⁵⁴ PetaBencana, ‘Software—PetaBencana.Id’ <<https://info.petabencana.id/research/software-2/>>.

⁵⁵ Rolnik (n 42).

⁵⁶ Australian Bureau of Statistics, ‘Twenty Years of Population Change’ (17 December 2020) <www.abs.gov.au/articles/twenty-years-population-change>.

⁵⁷ Eurostat, ‘Population Structure and Ageing: Statistics Explained’ (*Eurostat: Statistics Explained*, June 2021) <https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_structure_and_aging>.

⁵⁸ US Census Bureau, ‘65 and Older Population Grows Rapidly as Baby Boomers Age CB20-99’ (*Census.gov*, 25 June 2020) <www.census.gov/newsroom/press-releases/2020/65-older-population-grows.html>.

⁵⁹ Xizhe Peng, ‘Coping with Population Ageing in Mainland China’ (2021) 17 *Asian Population Studies* 1.

⁶⁰ General Comment No 4, para 8(e).

without direct or immediate input from users, can make both home and life in general safer; it will extend the time older persons can remain in their homes before requiring supported housing.⁶¹ This responds importantly to the strong link between the right to housing and the concept of home. While the right to housing does not protect a home per se, the seven elements of the right combine to protect housing that is freely chosen, individually meaningful, and connected to the local community (eg through the elements of location and cultural adequacy). Similarly, people with disabilities that require constant supervision or supportive assistance may be able to avoid having to live in care, in line with the right to independent living under article 19 of the Convention on the Rights of Persons with Disabilities (CRPD).⁶²

We recognise much of this technology is nascent—one recent meta-analysis finding no ‘convincing proof of a clear effect’⁶³ for ambient health monitoring. Nonetheless, the enormous financial and economic incentives for reducing the costs of aged and supportive care indicate the likelihood of continuing advancement in this sphere and may lead in important ways to the fulfilment of the right to housing as a human right. It is particularly important in recognising that states have an obligation to ensure that housing is accessible for groups with special needs or who are vulnerable such as older persons and persons with disabilities,⁶⁴ and that many situations of supported housing or institutionalised care can lead to a range of human rights violations.⁶⁵

In this section, we discuss a proposed platform for ambient assisted living for people with dementia in the early stages of cognitive decline. The AnAbEL platform is our exemplar, as it seeks to deal with shortcomings in existing ambient technology.⁶⁶ The platform seeks to meet two needs. It helps people with dementia maintain their independence for as long as possible. It also supports caregivers, as families may need to have someone take time away from work to support older family members. This technology will potentially reduce their need to be absent from the workplace while ensuring the safety of those they care for.

⁶¹ SJ Czaja and M Ceruso, ‘The Promise of Artificial Intelligence in Supporting an Aging Population’ (2022) *Journal of Cognitive Engineering and Decision Making* 16(4), 182–93. .

⁶² UN General Assembly (UNGA), Convention on the Rights of Persons with Disabilities (adopted by the General Assembly, 24 January 2007) UN Doc A/RES/61/106 (CRPD).

⁶³ Ju Wang and others, ‘Unobtrusive Health Monitoring in Private Spaces: The Smart Home’ (2021) 21 *Sensors* 864, 863.

⁶⁴ General Comment No 4, para 8(e).

⁶⁵ Eg as revealed by the Australian Royal Commission into Aged Care Quality and Safety. See Royal Commission into Aged Care Quality and Safety, ‘Final Report’ (26 February 2021) <<https://agedcare.royalcommission.gov.au/publications/final-report>>. See also Linda Steele and others, ‘Human Rights and the Confinement of People Living with Dementia in Care Homes’ (2020) 22(1) *Health and Human Rights Journal* 7.

⁶⁶ The platform was designed by scholars from the Research Group on Development of Intelligent Environments from the Department of Computer Science and the Centre for Public Health and Risk at Middlesex University.

The proposed system focuses on key safety domains: eating, drinking, sleeping, and bathing. The operating system draws information from several sensor systems throughout the house: an infrared movement detection system, detection of the use of electrical devices, sensors that identify if doors (eg cupboards and fridges) are open, if lights are on or off and pressure sensors to determine if a bed, sofa, or chair is being used. GIS information from the user's mobile is also imported into the system. After being provided with information about user timetables, the system then uses information gathered from ambient sensors built into the house to make assessments about when it needs to alert, in the first instance, the user, and when to escalate to the caregiver. Interaction with users is via mobile phone. The system needs to make multiple sets of decisions. In the first, it needs to interpret what input data means. Eating, for example, needs to be inferred from information pulled from the above sensors—perhaps a sequence of door openings, a presence in the kitchen, and sitting in a particular place. Second, the system needs to decide what departures from routine are usual. For example, does going to bed later than usual amount to an event of concern? Leaving the house, a task that might be fine at 11am, might be problematic at midnight.

There are other, similar systems under development and in use. For example, there is research on the use of wearable devices that capture user gestures, which could help persons with limited movement to interact with the technologies in their homes.⁶⁷ Another 'outside-the-house' example helps people with cognitive disabilities to access and navigate a transport system.⁶⁸ Presenting a map with an array of transport options, the device suggests a route using public transport based on GPS data. If the user departs from their proposed route—catching the wrong bus, for example—there is a machine-informed intervention. This technology might not appear immediately connected to the realisation of the right to housing occurring as it does outside the home. However, it points to the fact that realisation of the right to housing requires a home located where there are adequate and accessible transport links, community services, and opportunities to leave the home and move into the public sphere beyond.⁶⁹ It also points to the interconnections between all rights, not only economic and social (such as the right to work and to social security) but to civil and political rights such as freedom of movement and association.

⁶⁷ Amritha Purushothaman and Suja Palaniswamy, 'Development of Smart Home Using Gesture Recognition for Elderly and Disabled' (2020) 17 Journal of Computational and Theoretical Nanoscience 177.

⁶⁸ S Carmien and others, 'Socio-Technical Environments Supporting People with Cognitive Disabilities Using Public Transportation' (2005) 2 ACM Transactions on Computer-Human Interaction 30.

⁶⁹ On the location of the right to housing at the intersection of the public and private, see also Jessie Hohmann, 'Conceptualising Domestic Servitude as a Violation of the Human Right to Housing and Reframing Australian Policy Responses' (2022) 31(1) Griffith Law Review 98; Hohmann (n 8) 145–65.

There are, as with all technologies, concerns. Isolation is a major problem for any person who is older or has a debilitating disability.⁷⁰ There is an obvious temptation in a world of cost-cutting, economic rationalism, and tech-centric interventions that the very human need for meaningful emotional contact will be disregarded. While remaining in place or asserting the degree of independence that one wants supports the right to housing, the need for meaningful social integration is also related to the right. The right is not realised if people are ghettoised or are unable to experience both the public and private sphere. In fact, housing—and rights to and in it—sit at the crucial juncture of the public and the private, and the right will not be realised if people are confined to their housing against their will, no matter how adequate that housing otherwise is.⁷¹ Striking a balance between these countervailing objectives must support the objectives of individuals involved, including residents and caregivers, and not driven by class-based opportunity or funding related-imperatives.

4.3 Sanitation

The availability of services, including sanitation, is an inherent aspect of the right to adequate housing.⁷² Failures of sanitation extend far beyond simple access to a toilet or latrine. The World Health Organization (WHO) reports that, as of 2017, only 45 per cent of the global population use a ‘safely managed’ sanitation service, and two billion people do not have access to even basic toileting facilities, with over 670 million people still defecating openly (ie in public—gutters, bushes, beaches, or the like).⁷³ Residents of informal settlements are particularly likely to lack access to adequate sanitation facilities.⁷⁴ While connection to a sewerage system remains gold-standard, attaining this goal involves a high cost, challenges around land tenure, and political considerations.⁷⁵ This means that sanitation issues at the informal settlement level are often more effectively dealt with at the household or community level, and often entail regular emptying of on-site reservoirs, sometimes on a daily basis.⁷⁶

⁷⁰ Oliver Hämmig, ‘Health Risks Associated with Social Isolation in General and in Young, Middle and Old Age’ (2019) 14 PLoS One e0219663; Stephen J Macdonald and others, ‘“The Invisible Enemy”: Disability, Loneliness and Isolation’ (2018) 33 Disability & Society 1138.

⁷¹ Hohmann (n 8) 145–65.

⁷² UN CESCR, ‘General Comment No 15: The Right to Water (Arts 11 and 12 of the Covenant)’ (20 January 2003) UN Doc E/C.12/2002/11. An implied right to water and sanitation arises from arts 11(1) and 12 of the ICESCR.

⁷³ World Health Organization (WHO), ‘Sanitation Fact Sheet’ <www.who.int/news-room/fact-sheets/detail/sanitation>.

⁷⁴ Sheela S Sinharoy, Rachel Pittluck, and Thomas Clasen, ‘Review of Drivers and Barriers of Water and Sanitation Policies for Urban Informal Settlements in Low-Income and Middle-Income Countries’ (2019) 60 Utilities Policy 100957.

⁷⁵ *ibid.*

⁷⁶ *ibid.*

Because waste collection is expensive and time-consuming, there are significant incentives to devolve this responsibility to the individual household (eg using a septic system) or eschewing a regularly scheduled pick-up service in favour of ‘just-in-time’ latrine emptying. Here, we discuss a machine learning (ML) system that is trained using historical use data and weight sensors in order to estimate how full latrines are. This information was used to predict latrine overflow events and thus to facilitate dynamic scheduling. When tested by a private sanitation collection service in an informal settlement in Nairobi, the system was useful for managing lower-use latrines and able to predict overflow events with a high degree of accuracy.⁷⁷ Because many users of container-based waste services pay privately for the collection of their waste,⁷⁸ and poverty has a direct impact on access to and use of sanitation services,⁷⁹ even small improvements in collection efficacy have a potential impact on sanitation outcomes. On the other hand, such private responsibility should not detract from state responsibility toward those in informal settlements for appropriate, adequate facilities (as required by the UN CESCR’s General Comment No 4),⁸⁰ to ensure that people can access adequate housing—that is, housing that in the terms of the right does not endanger the safety of the residents including through poor sanitation, that has all appropriate facilities, and that is not built on polluted sites (including sites polluted by effluent).

Monitoring the cleanliness of shared toileting facilities can be an expensive, labour-intensive exercise, with ‘a huge amount of money and manpower to maintain these public toilets’.⁸¹ Here, we draw the reader’s attention to potential applications of AI as a mode of remedy. In one project, for example, sensors were used as an ‘e-nose’ to detect unpleasant odours—that is, bad smells associated with the malfunction or uncleanliness of the toilet facility.⁸² The challenge presented to the system is differentiating between different types of odour, while correcting for the impact that humidity levels have on sensor receptivity. While still in the experimental phase, the system can differentiate between excreta types and other smells and has the potential to reduce the costs of sanitation monitoring in hard to reach places while simultaneously triggering cleaning processes when required.

⁷⁷ Nick Turman-Bryant and others, ‘Toilet Alarms: A Novel Application of Latrine Sensors and Machine Learning for Optimizing Sanitation Services in Informal Settlements’ (2020) 5 Development Engineering 100052.

⁷⁸ Caroline Jennings Saul and Heiko Gebauer, ‘Digital Transformation as an Enabler for Advanced Services in the Sanitation Sector’ (2018) 10 Sustainability 752.

⁷⁹ Christophe Bosch and others, ‘Water, Sanitation and Poverty’ [2001] Draft chapter. Washington DC: World Bank.

⁸⁰ General Comment No 4, para 8(b).

⁸¹ Prasad Deshmukh and others, ‘Intelligent Public Toilet Monitoring System Using IoT’ (IEEE, 2020) <<https://ieeexplore.ieee.org/document/9297839>>.

⁸² Jin Zhou and others, ‘Sensor-Array Optimization Based on Time-Series Data Analytics for Sanitation-Related Malodor Detection’ (2020) 14 IEEE Transactions on Biomedical Circuits and Systems 705.

The commonality between the three potential uses of AI-informed technologies and insights vis-à-vis the right to housing focus is the mobilisation of individual or community resources in the face of, or against, an apathetic state. While we support any increase in the capacity of the ability of individuals to assert rights, we note they reflect an expectation that individuals need to actively claim their rights, often at their own expense. The reification of resilience has an increasing centrality to disaster recovery and climate adaptation.⁸³ The cultivation of actors that are ‘confident and at home with contingency and the unexpected’⁸⁴ alleviates pressure on the state and non-governmental actors to facilitate rights, including the right to housing. These tools should not be another opportunity for the state to revile from its obligations in favour of profit-seeking corporations, with little regard for the enjoyment of adequate housing by ordinary people.

5 Conclusion

The use of Big Data and AI can have both problematic and positive impacts on the right to housing. The technologies’ uses will be applied on top of already existing power relations, biases, and structural inequalities. It is power imbalances between tenants and landlords, the propertyless and the propertied, the enabled and the disabled, that create opportunities and potential for discrimination. Power relations may not change, only be ‘masked in new ways’ by these technologies.⁸⁵ The *novelty* of the risk of new technologies in the field of housing is often also overstated; as McElroy argues, property itself has long operated as a technology of discrimination and dispossession.⁸⁶

However, due to the structural inequalities and power imbalances that do exist, the application of AI in the field of housing does present serious challenges for the right to housing. So, for example, while technologies of the smart home have powerful potential to enable people to live in their homes for longer, with greater independence, safety, and security, we must be attuned to the potential that these technologies will be used in those protective ways for First World, middle-class homeowners, while they are simultaneously used to surveil and discriminate against poor and racialised people in their access to housing. At the same time, disaster risk reduction technologies can be used to respond to unfolding disasters

⁸³ Mark Duffield, ‘The Resilience of the Ruins: Towards a Critique of Digital Humanitarianism’ (2016) 4 *Resilience* 147; Mark Duffield, *Post-Humanitarianism: Governing Precarity in the Digital World* (Wiley 2018).

⁸⁴ David Chandler, *Ontopolitics in the Anthropocene: An Introduction to Mapping, Sensing and Hacking* (1st edn, Routledge 2018) 142 <www.taylorfrancis.com/books/9781351335928>.

⁸⁵ Shaw (n 24) 1054.

⁸⁶ See also Erin McElroy, ‘Property as Technology: Temporal Entanglements of Race, Space, and Displacement’ (2020) 24 *City* 112; see also Brenna Bhandar, *Colonial Lives of Property: Law, Land, and Racial Regimes of Ownership* (Duke UP 2018).

to minimise harm to people (and their homes) and to prepare areas to be more robust and protected, they might also more likely be used to map, see, and know areas from which additional value can be extracted by the financially powerful, at the expense of the homes and security of existing populations, for whom the technologies just become another tool for eviction and relocation.

Given the array of opportunities and potential threats presented by the use of AI and other technologies, it is important to coalesce around an ideological stance to navigate the various options. We argue that, when it comes to decisions about directions that technologies take us, human rights are an important standard against which we can measure the acceptability of technological change. With respect to algorithmic systems, for example, human rights, including the ‘rich discourse and practice’ of economic, social, and cultural rights,⁸⁷ we echo recent scholarship in stressing that HRL ‘offers an organising framework for assessing algorithms, bringing the language of law and human rights back to the fore (instead of loosely used terms such as bias, harm) and emphasising the obligations of states’.⁸⁸ While human rights cannot be the only response to AI, importantly, human rights provide principles against which an evaluation can be made, and a common language and set of objectives to guide positive, human-focused change.

⁸⁷ Jędrzej Niklas, ‘Conceptualizing Socio-Economic Rights in the Discussion on Artificial Intelligence’ (SSRN, 26 April 2019) <<https://ssrn.com/abstract=3569780>> or <<http://dx.doi.org/10.2139/ssrn.3569780>>, 3.

⁸⁸ ibid; Rachovitsa and Johann (n 32) 8, 15.

25

Artificial Intelligence and Human Rights at Work

Joe Atkinson and Philippa Collins

1 Introduction

Artificial intelligence (AI) is disrupting and transforming our lives across multiple dimensions of society. AI is now used to make decisions previously undertaken by humans in contexts such as policing, social security, immigration, and in the workplace. While the use of AI to automate work processes might eventually lead to a level of worker displacement and job destruction that threatens to undermine the right to work,¹ it is clear that the use of AI to manage and govern the workplace presents the more immediate and pressing challenge. This chapter addresses this latter innovation, namely the impact of technology on the *qualitative* rather than *quantitative* dimension of the future of work. Specifically, it is concerned with the implications for human rights of what Mateescu and Nguyen describe as ‘algorithmic management’, that is, a ‘diverse set of technological tools and techniques to remotely manage workforces, relying on data collection and surveillance of workers to enable automated or semi-automated decision-making’.²

Our argument here is that the rise of algorithmic management poses a significant and pervasive threat to human rights at work, one that is not confined to the rights of privacy and equality concerns that have so far dominated scholarly attention.³ The use of these tools has the potential to frustrate the protection of workers’ human rights, which is an important normative goal for labour law. In section 2, we

¹ For discussion of this prospect and possible legal responses, see Cynthia Estlund, *Automation Anxiety: Why and How to Save Work* (OUP 2021).

² Alexandra Mateescu and Aiha Nguyen, ‘Algorithmic Management in the Workplace’ (2019) Data & Society 1.

³ See eg Robert Sprague, ‘Welcome to the Machine: Privacy and Workplace Implications of Predictive Analytics’ (2014) 21 Richmond Journal of Law and Technology 1; Bart Custers and Helena Ursic, ‘Workers’ Privacy in a Digitalized World under European Law’ (2018) 39 Comparative Labor Law and Policy Journal 323; Ifeoma Ajunwa, Kate Crawford, and Jason Schultz, ‘Limitless Worker Surveillance’ (2017) 105 California Law Review 735; Ifeoma Ajunwa, ‘Algorithms at Work: Productivity Monitoring Applications and Wearable Technology as the New Data-Centric Research Agenda for Employment and Labor Law’ (2018) 63 St Louis University Law Journal 21; Jeremias Adams-Prassl, ‘What if Your Boss Was an Algorithm? Economic Incentives, Legal Challenges, and the Rise of Artificial Intelligence at Work’ (2019) 41 Comparative Labor Law and Policy Journal 123.

elaborate upon the phenomenon of algorithmic management and illustrate how the integration of AI, and the data collection that underpins it, affects the working lives of individuals. In the subsequent part, we draw on the existing literature that paints a clear picture of how privacy and data protection rights, as well as the right to equality, are threatened by the deployment of algorithmic management processes. We go beyond these current analyses, however, by highlighting how algorithmic management poses a broader threat to human rights at work. Algorithmic management entails risks to a wide range of workers' rights in addition to privacy and equality, such as freedom of association, expression, thought, and belief, as well as due process rights and rights to decent working conditions.

In the final section, we argue that an adequate response to the risk to human rights presented by algorithmic management must involve the use of *ex ante* methods of regulation rather than merely relying on *ex post* responses and litigation. We identify and discuss two important pre-emptive means of ensuring employers' choices regarding algorithmic management respect the human rights of workers. First, collective bargaining over the use of these technologies, and second, legal duties to conduct a robust assessment of any human rights impacts before implementing a new policy or data processing method. Both these forms of *ex ante* governance have the potential to place limits upon algorithmic management that are tailored to the particular organisational context in question and ensure that workers are protected from excessive monitoring and unjust data-driven practices.

2 Algorithmic Management and the Digital Revolution at Work

Algorithmic management is deployed at each major point of contact between a worker and their employer. From recruitment of new staff and the day-to-day management of tasks, through to disciplinary action and the termination of employment, management functions can now be delegated entirely to technology or significantly bolstered by the use of algorithms.⁴ Whilst algorithmic management is currently most prevalent in platform work and the 'gig economy', these practices are rapidly spreading to other sectors of the labour market.⁵ This shift has recently been accelerated by the COVID-19 pandemic and a shared desire amongst many employers to use technology to manage a newly remote workforce.⁶

⁴ Trades Union Congress (TUC), 'Technology Managing People: The Worker Experience' (2020) <www.tuc.org.uk/sites/default/files/2020-11/Technology_Managing_People_Report_2020_AW_Optimised.pdf>.

⁵ Alex J Wood, *Algorithmic Management: Consequences for Work Organisation and Working Conditions* (JRC Working Papers Series on Labour, Education and Technology 2021/07; European Commission 2021) 1.

⁶ See Abigail Gilbert and Anna Thomas, *The Amazonian Era: How Algorithmic Systems Are Eroding Good Work* (Institute for the Future of Work 2021) 1, 17.

The first point of contact with an employer, where a candidate is applying for a job, is the most likely to be mediated by an algorithm.⁷ It is increasingly common that seeking out and shortlisting applicants is conducted by algorithm. Using a set of training data, including online information, written answers, or video interviews, algorithms can identify features of desirable candidates for the position. Some such features may be obvious, such as mentioning a particular technical skill or experience, but others may be apparently random factors that happen to be prevalent amongst successful candidates. In this way, candidates' application materials can be screened and sifted by AI software, which continues to learn and adapt as it is exposed to more data. AI software programs can also be used during hiring to scrutinise social media profiles of potential employees for desirable, or more likely *undesirable*, characteristics that are revealed by their posts. Finally, video recordings of candidates answering pre-set questions can be assessed by AI tools, such as HireVue, to generate an 'employability score' based on any number of visual, verbal, or behavioural traits of the applicant. The complexity and opacity of these AI-based hiring processes adds to the difficulties that candidates for roles already face in understanding how decisions that affect their livelihood are made.

Once the candidate is appointed, the potential for algorithmic management continues. As Alex Wood observes, algorithmic management can be seen across a variety of employer functions, particularly in the direction, evaluation, and discipline of workers.⁸ In terms of the direction of workers, a common use of algorithms is in the creation and allocation of shift patterns based on predictions of future demand. For example, Percolata's program uses predictive analytics to create a schedule that aims to maximise sales. AI generates recommendations of the optimal mix of workers, and their allocated tasks, for every fifteen-minute time period throughout the day. Such programs enable employers to match the amount of labour contracted for (and the associated costs of this) precisely with expected demand, which facilitates and incentivises the use of atypical working arrangements thus contributing to the continued fragmentation and fissuring of workplaces.⁹

The allocation of work between workers and the pace of the work to be performed is also frequently determined by algorithmic tools. The delivery driver with the nearest GPS location is sent to pick up the takeaway food, whilst pickers in the Amazon fulfilment centres must adhere to an algorithmically determined 'Amazon pace' along the route to the next item they need to pick: not running but walking as fast as possible.¹⁰ Workers may also receive directions on how to complete their

⁷ TUC (n 4).

⁸ Wood (n 5).

⁹ Judy Fudge, 'Fragmenting Work and Fragmenting Organizations: The Contract of Employment and the Scope of Labour Regulation' (2006) 44 Osgoode Hall Law Journal 609; David Weil, *The Fissured Workplace* (Harvard UP 2014).

¹⁰ Alessandro Delfanti, 'Machinic Dispossession and Augmented Despotism: Digital Work in an Amazon Warehouse' (2021) 23 New Media & Society 39, 47.

allocated tasks via a mobile application or wearable device. Whilst previously workers would require some training and understanding of the processes in their workplace, the co-ordination of work can now be done by algorithm. Complex processes are divided into ever smaller and simpler components, so that each task can be completed with minimal training—workers need only follow pictorial directions on a handheld device in order to complete their tasks.¹¹

Once work is allocated, AI technologies can be deployed to monitor and evaluate the performance of tasks on a moment-to-moment basis using data points gleaned from numerous sources, with the goal of optimising the efficiency of outputs or performance metrics identified by the organisation. Data is then processed and analysed so that it can be acted upon either by management or by the worker themselves, as they respond to real-time corrections and recommendations about their work performance. For example, a program used by call centres, Cogito, engages in real-time voice evaluation, providing prompts to workers during calls based on this analysis such as to be more empathetic or to talk more slowly.¹² Increasingly, businesses are using client and customer ratings as an important source of information for algorithmic assessments of an individual's performance. Platforms that offer services online, such as data entry, translation, and programming, combine customer ratings with digital monitoring based on keystrokes and screenshots to evaluate workers.¹³ Keystroke logging, screenshots, and automated analysis of electronic communications are also used to monitor and evaluate those working remotely and at home. The shift away from periodic performance reviews to continuous and instantaneous evaluation is underpinned by 'prolific data collection and surveillance of workers through technology'.¹⁴ Such pervasive monitoring has significant negative consequences for workers, such as being subject to heightened levels of subordination and control, the intensification of work processes, and increased risks to occupational health and safety.¹⁵

Just as algorithms and technology play a role in the other stages of an employment relationship, so too do they in the discipline and termination of the relationship. Software can be used to identify problems with a worker's attendance or performance, and either flag this to the employer or, less frequently, implement disciplinary action directly. At Uber, for instance, algorithms evaluate workers on

¹¹ Simon Schaupp, 'Algorithmic Integration and Precarious (Dis)Obedience: On the Co-Constitution of Migration Regime and Workplace Regime in Digitalised Manufacturing and Logistics' (2021) 36 *Work, Employment and Society* 310, 317.

¹² Kevin Roose, 'A Machine May Not Take Your Job, but One Could Become Your Boss' *New York Times* (23 June 2019).

¹³ See Alex J Wood and others, 'Good Gig, Bad Gig: Autonomy and Algorithmic Control in the Global Gig Economy' (2019) 33 *Work, Employment and Society* 56.

¹⁴ Mateescu and Nguyen (n 2).

¹⁵ Gilbert and Thomas (n 6); Karolien Lenaerts and others, *Digital Platform Work and Occupational Safety and Health: A Review* (European Agency for Safety and Health at Work 2021); Phoebe Moore, *OSH and the Future of Work: Benefits and Risks of Artificial Intelligence Tools in Workplaces* (European Agency for Safety and Health at Work 2019); Wood (n 5).

the basis of their customer rating and the rate at which they accept jobs offered to them on the app. If an individual's score falls below the level deemed acceptable, the individual will be temporarily removed from the app as a disciplinary measure, meaning they will not have access to work for that period of time. If the driver's score is consistently below the algorithm's expectation or if the AI software flags fraudulent behaviour, the individual will be removed from the app permanently.¹⁶ Similarly, Amazon's system monitors each workers' productivity and can issue warnings and automatic terminations if their productivity is not high enough.¹⁷ In other systems, workers with lower productivity scores may receive less work or find their ability to book into shifts restricted,¹⁸ meaning that it may become untenable to continue working through that platform or employer. Built upon a wider structure of digital monitoring and real-time surveillance of workers, the termination of employment based on an algorithm's recommendation is the natural endpoint of management-by-algorithm.

3 The Human Rights Dimensions of Algorithmic Management

The early years of analysing AI technologies in the workplace have been dominated by discussions of their implications for workers' privacy, data protection, and discrimination rights.¹⁹ The rise of algorithmic management undoubtedly threatens these rights. In terms of informational privacy and data protection, workers are subject to increased monitoring and surveillance across a wider range of data points, which may include personal data relating to their health or lives outside of work. It is difficult for workers to understand what information is being collected on them, how this is being used and shared with others, as well as the risks involved in these processes. Indeed, in many instances, workers are not even aware of the technologies being used to recruit or manage them.²⁰ In some cases, employers incentivise or demand the sharing of intimate data with the organisation through a repurposing of smart wristwatches or similar wearable devices which track the

¹⁶ Alex Rosenblat and Luke Stark, 'Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers' (2016) 10 International Journal of Communication 3758, 3774–75; Sarah Butler, 'Court Tells Uber to Reinstate Five UK Drivers Sacked by Automated Process' *Guardian* (14 April 2021); and Worker Info Exchange, *Managed by Bots: Data-Driven Exploitation in the Gig Economy* (2021) <www.workerinfoexchange.org/wie-report-managed-by-bots>.

¹⁷ Colin Lecher, 'How Amazon Automatically Tracks and Fires Warehouse Workers for "Productivity"' (*The Verge*, 25 April 2019) <www.theverge.com/2019/4/25/18516004/amazon-warehouse-fulfillment-centers-productivity-firing-terminations> and Spencer Soper, 'Fired by Bot at Amazon: "It's You Against the Machine"' (*Bloomberg*, 28 June 2021) <www.bloomberg.com/news/features/2021-06-28/fired-by-bot-amazon-turns-to-machine-managers-and-workers-are-losing-out>.

¹⁸ Wood (n 5) 8.

¹⁹ See n 3.

²⁰ TUC (n 4).

wearer's location, heart rate, sleep, and activity rates.²¹ Automated social media monitoring tools also track and monitor worker behaviour and speech away from the workplace, and this information may be used as the basis for disciplinary action.²² The right to informational and behavioural privacy is therefore jeopardised, not only during working time but also beyond.

In terms of discrimination and equality, the technologies used to select, direct and evaluate workers outlined above serve to replicate, entrench, and accentuate existing workplace inequalities.²³ The risk of bias in algorithmic decision-making is now firmly established in the research and recognised by policymakers.²⁴ While it will (hopefully) be rare for algorithmic models to incorporate protected characteristics such as race, religion, or gender directly, algorithmic decision-making may be discriminatory by relying on combinations of other factors that amount to close proxies to these characteristics. The combination of postcode and educational history, for instance, may act as a proxy for ethnicity in some circumstances.²⁵ In addition to relying on protected characteristics or their closely correlated data points, algorithmic management may give rise to discrimination where the models reflect the biased assumptions and choices of the programmers, or where a machine learning model is developed using 'training data' that contains bias or historical discrimination. Such algorithms are likely to reproduce, and even amplify, existing inequalities and historic discrimination in the workplace.

In the context of recruitment, for example, Kelly-Lyth cites examples such as an Amazon algorithm that was abandoned after marking down applications that contained the word 'women's' (as in women's sports teams or colleges), and others that have learnt to associate female names with domestic duties.²⁶ Facial recognition and analysis algorithms have also been shown to discriminate against people of colour and pose challenges for people with disabilities that affect their facial

²¹ See Philippa Collins and Stefania Marassi, 'Is That Lawful? Data Privacy and Fitness Trackers in the Workplace' (2021) 37 International Journal of Comparative Labour Law and Industrial Relations 65.

²² Lisa Kresge, 'Data and Algorithms in the Workplace: A Primer on New Technologies' (Working Paper UC Berkeley Labor Center Technology and Work Program, 2020) 6 <<https://laborcenter.berkeley.edu/wp-content/uploads/2020/12/Working-Paper-Data-and-Algorithms-in-the-Workplace-A-Primer-on-New-Technologies-FINAL.pdf>>.

²³ Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books 2016); Sandra G Mayson, 'Bias In, Bias Out' (2019) 128 Yale Law Journal 2218; and Pauline T Kim, 'Data-Driven Discrimination at Work' (2017) 58 William & Mary Law Review 857. For analysis of how UK discrimination law applies in this context, see Joe Atkinson, 'Automated Management and Liability for Digital Discrimination under the Equality Act 2010' (*UK Labour Law Blog*, 2020) <<https://uklabourlawblog.com/2020/09/10/automated-management-and-liability-for-digital-discrimination-under-the-equality-act-2010-by-joe-atkinson/>>.

²⁴ See Centre for Data Ethics and Innovation, *Review into Bias in Algorithmic Decision-Making* (2020); and All-Party Parliamentary Group on the Future of Work, *The New Frontier: Artificial Intelligence at Work* (2021) 26.

²⁵ See Part IV, in particular the chapter by Louis Koen and Kgomo Mufamadi in this volume.

²⁶ Aislinn Kelly-Lyth, 'Challenging Biased Hiring Algorithms' (2021) 41 Oxford Journal of Legal Studies 889.

movements.²⁷ A person with a physical disability may well be disadvantaged by an algorithm that controls the pace of work, particularly as the employer's duty to make reasonable accommodations is unlikely to be reflected in the design of the software. Similarly, algorithms that allocate shifts and approve holiday requests are unlikely to take into account factors such as individuals' caring responsibilities or religious beliefs. A final concern in respect of discrimination is the role of customer or client evaluations in algorithmic management, as these may be tainted by the conscious or unconscious bias of the customer and lead to biased performance evaluations, with dramatic consequences for workers' livelihoods.²⁸

Whilst the above discussion illustrates that algorithmic management is a serious threat to workers' privacy and equality rights, we argue that the risk posed to human rights at work is much broader than has been appreciated thus far. As labour lawyers, we might start by considering the right to form and join a trade union for the protection of one's interests, contained in the right to freedom of association in article 11 of the European Convention on Human Rights (ECHR) and article 22 of the International Covenant on Civil and Political Rights (ICCPR). Algorithms can be used by employers to 'get ahead' of workers' attempts to unionise, targeting their efforts to ensure that workers do not exercise their right to freedom of association. In 2020, Wholefoods in the United States (US) was revealed to use a 'heatmap' to predict which stores were at risk of unionisation based upon a combination of metrics. This algorithmic prediction was calculated using a combination of external risks (size and proximity of local unions to the store, local unemployment rate, rate of union-related incidents and complaints to the National Labor Relations Board), store risks (stores with lower racial and ethnic diversity and lower wage rates were flagged as higher risk for unionisation) and team member sentiment, drawn from surveys of employees.²⁹ Employers can also use technological monitoring and analysis of worker interactions to identify individuals likely to be involved in unionisation efforts and then take steps to try to prevent this from happening. For example, software can be used to scan workers' emails, personal messenger communications, or conversations recorded via wearable devices for key words and phrases relating to union activities. Similarly, Google has introduced a tool that flags and

²⁷ Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) Proceedings of Machine Learning Research: Conference on Fairness, Accountability and Transparency 1–15; Kerri A Thompson, 'Countenancing Employment Discrimination: Facial Recognition in Background Checks' (2020) 8 Texas A&M Law Review 63.

²⁸ See Alex Rosenblat and others, 'Discriminating Tastes: Customer Ratings as Vehicles for Bias' (2016) *Data & Society* 1.

²⁹ Jay Peters, 'Whole Foods is Reportedly Using a Heat Map to Track Stores at Risk of Unionization' (*The Verge*, 20 April 2020) <www.theverge.com/2020/4/20/21228324/amazon-whole-foods-unionization-heat-map-union>.

monitors large internal meetings of employees which clearly has the potential to be used to identify and suppress union organising.³⁰

Less directly, the working environment created through algorithmic management makes it difficult for workers to exercise their rights to join trade unions and act collectively that are protected as part of freedom of association. Algorithmic management enables increased use of zero-hours contractors, agency workers, and other forms of precarious work where there are significant hurdles to collective organising. Similarly, constantly changing shift patterns determined by algorithms, and the digital enforcement of a high pace of work during shifts, leave workers with little time or opportunity to develop the solidarity and social bonds needed for unionisation and the exercise of collective power within the workplace.³¹ To give a specific example, Deliveroo has been held to discriminate against trade union members participating in strike action (amongst other groups) by an Italian labour court.³² Deliveroo's algorithm removed union members from a 'priority group' who enjoyed privileges with regard to work allocation because they failed to attend a booked slot. By failing to take into account *why* the rider had not attended, namely because they were taking industrial action, Deliveroo discriminated against trade union members and the company was ordered to correct the discrimination and pay compensation.

Management by algorithm can also stymie worker attempts to exercise voice and negotiate collectively over their terms and conditions of work. The deployment of data-driven management strategies widens the existing information asymmetries between workers and employers.³³ Thus, any bargaining efforts commence from a more unequal starting position. Moreover, employers will not generally themselves be in control of the design of AI systems they deploy: these systems are frequently 'bought in' from a specialist company that retains proprietary rights over the process and may refuse to share information from within the 'black box'. In such situations there are additional hurdles to unions and employers seeking to negotiate, in that any agreement about the internal operation of an algorithm will not have any effect unless a third-party supplier puts it into practice. This they may refuse to do, leaving the employer little choice (one would hope) but to stop implementation of the proposed strategy. Once introduced, algorithmic management is also not conducive to the input and influence on individual decisions that union representatives may otherwise have. These systems are less adaptive to individual

³⁰ Nick Statt, 'Google Accused of Spying with New Tool that Flags Large Employee Meetings' (*The Verge*, 23 October 2019) <www.theverge.com/2019/10/23/20929524/google-surveillance-tool-accused-employee-activism-protests-union-organizing>.

³¹ John Holland, 'Amazon Inquiry' (*Notes From Below*, 7 October 2020) <<https://notesfrombelow.org/article/amazon-inquiry>>.

³² Tribunal of Bologna, Order no 2949/2019 (31 December 2020). See commentary in Vincenzo Pietrogiovanni, 'Deliveroo and Riders' Strikes: Discriminations in the Age of Algorithms' (2021) 7 International Labour Rights Case Law 317.

³³ Rosenblat and Stark (n 16).

circumstances and have less room for discretion or ability to take workers' interests into account than human decision-makers. In addition, the data points used by the algorithm and weight ascribed to them are likely to be entirely hidden from view. Worker representatives will, therefore, struggle to understand or have meaningful influence over the systems by which their work and livelihood is managed.

Away from freedom of association, algorithmic management threatens a range of other human rights at work. Automated monitoring of a worker's emails or social media posts, without more, can constitute an interference with the right to freedom of expression, which would only be compounded if their speech is flagged as in breach of company policies and leads to disciplinary action against the individual.³⁴ Where workers are made aware of any such monitoring practices, as is required by the European Court of Human Rights (ECtHR),³⁵ this would have a chilling effect on their expression and cause employees to be inhibited in their interactions within and beyond the workplace. Emotion recognition and sentiment analysis technologies, which seek to 'read' what an individual is feeling and thinking from audio or visual data, will also infringe upon a person's ability to form their own views freely and without being penalised.³⁶ Attempts to glean information about these most intimate aspects of a person's thoughts and reactions are a source of significant concern from the perspective of the right to freedom of thought, conscience, and belief.³⁷

The rights to due process at work, of the kind that are protected under article 6 and the procedural aspects of the other ECHR rights,³⁸ are also at risk where staff are managed with minimal human intervention. The contracts of Uber drivers, for example, are terminated where the AI system detects what it understands as fraudulent behaviour.³⁹ Although Uber states on its website that there is human review of the flagged behaviour before termination,⁴⁰ these workers have no meaningful opportunity to influence or challenge these technology-driven decisions to

³⁴ Virginia Mantouvalou, 'I Lost My Job over a Facebook Post: Was That Fair?' Discipline and Dismissal for Social Media Activity' (2019) 35 International Journal of Comparative Labour Law and Industrial Relations 101; *Bărbulescu v Romania* App no 61496/08 (ECtHR, 5 September 2017); *Antović and Mirković v Montenegro* App no 70838/13 (ECtHR, 28 November 2017).

³⁵ *Bărbulescu* (n 34) discussed in Joe Atkinson, 'Workplace Monitoring and the Right to Private Life at Work' (2018) 81 Modern Law Review 688.

³⁶ Discussed in Valerio De Stefano, 'Neurosurveillance and the Right to Be Human at Work' (*On Labor*, 15 February 2020) <<https://onlabor.org/neuro-surveillance-and-the-right-to-be-humans-at-work/>>.

³⁷ See the chapter by Jeroen Temperman in this volume.

³⁸ Philippa Collins, *Putting Human Rights to Work: Labour Law, the ECHR and the Employment Relation* (OUP 2022) 67–68.

³⁹ Natasha Bernal, 'They Claim Uber's Algorithm Fired Them. Now They're Taking it to Court' (*Wired*, 2 November 2020) <www.wired.co.uk/article/uber-fired-algorithm>.

⁴⁰ Uber, 'Fraud Activities on the Uber Driver App' <www.uber.com/gb/en/drive/driver-app/fraud-activities>.

flag their conduct and end their employment—or even to understand why this has happened.⁴¹

Finally, the heightened intensity of work that results from algorithmic management practices will frequently harm worker's physical and mental health, thereby threatening their rights to health and bodily security. There is mounting evidence that use of workplace surveillance and algorithmic management tools leads to high levels of stress and creates significant risks for occupational health and safety.⁴² For example, an app used in engineering settings monitors how quickly every worker completes particular tasks in order to find the fastest operator. The app then calculates 95 per cent optimisation for that task, in relation to the quickest worker, and all staff are expected to comply with that work rate.⁴³ Such a pace of work may not be achievable for every worker, leading them to push themselves physically to meet the demands set by the AI management software and thereby jeopardising the right to a healthy and safe work environment.⁴⁴

Taking a step back, we can see that the deployment of AI technologies to manage workforces places a downwards pressure on the quality of work and threatens to undermine the right to fair and just working conditions found in article 7 of the International Covenant on Economic, Social and Cultural Rights (ICESCR). Individuals working 'in the shadow' of an algorithmic boss feel constrained by the knowledge that they are subject to monitoring and that the data is used to determine their access to work or to mete out sanctions. As workers do not know the data points used in these technologies, they may attempt to predict what behaviour will be viewed favourably by the algorithm and engage in 'anticipatory compliance practices', thereby internalising its assumed decision-making processes.⁴⁵ This kind of anticipatory behaviour amounts to a general 'chilling effect' on a person's willingness to exercise their human rights freely,⁴⁶ as workers attempt to pacify the algorithm by refraining from exercising their rights such as freedom of expression or association in ways they believe might lead them to be penalised. While not a comprehensive survey, the above analysis demonstrates the pervasive threat that AI poses to human rights at work, one that extends beyond the rights to privacy and equality.

⁴¹ See Philippa Collins, 'Automated Dismissal Decisions, Data Protection and The Law of Unfair Dismissal' (*UK Labour Law Blog*, 2021) <<https://uklabourlawblog.com/2021/10/19/automated-dismissal-decisions-data-protection-and-the-law-of-unfair-dismissal-by-philippa-collins>>.

⁴² See n 15.

⁴³ Gilbert and Thomas (n 6) 13.

⁴⁴ See further Adrian Todoli-Signes, 'Making Algorithms Safe for Workers: Occupational Risks associated with Work Managed by Artificial Intelligence' (2021) 27 Transfer 433.

⁴⁵ Eliane Leontine Bucher, Peter Kalum Schou, and Matthias Waldkirch, 'Pacifying the Algorithm: Anticipatory Compliance in the Face of Algorithmic Management in the Gig Economy' (2021) 28 Organization 44, 52.

⁴⁶ For an example of surveillance technologies being found to infringe other rights due to this chilling effect, see *Big Brother Watch v United Kingdom* Application no 58170/13 (ECtHR, 25 May 2021).

4 Thinking Ahead: An Ex Ante Approach

Once the true extent of the threat to workers' fundamental rights generated by algorithmic management is understood, it is important to consider how the risk to these rights can be addressed and minimised. Although ex post legal frameworks with appropriate remedies are undoubtedly necessary to regulate the use of AI and protect rights in the workplace, the effectiveness of these measures in this fast-moving area of technology is limited due to their reactive nature. By the time litigation has made its way through the courts, or new legislation is introduced to address an identified harm, the practices of employers and issues faced by workers are likely to have evolved. Moreover, even where remedial regimes exist, it will be important for these to be supported by preventative policies and frameworks aimed at ensuring employers do not deploy workplace technologies in a manner incompatible with human rights.

In an area that is moving so rapidly it is particularly useful to consider methods of regulation that have the primary goal of preventing infringements of workers' human rights from the use of algorithmic management in advance of their occurrence. For example, the European Union's (EU) draft AI Act adopts an approach along these lines to 'high risk' systems, which includes software used to hire, select, manage, or terminate employment.⁴⁷ The draft Act imposes obligations upon the provider (ie the developer) of the software to ensure, in advance of marketing the system, that it meets requirements such as appropriate data and data governance systems, transparency, human oversight and accuracy, robustness, and cybersecurity.⁴⁸ These ex ante obligations are central to the Act, although they are supplemented by obligations to monitor the use and impacts of the system once it is deployed by the user. Whilst there are legitimate concerns about the reliance upon methods of self-assessment by providers,⁴⁹ this approach does have the benefit of seeking to prevent harms rather than merely providing a remedy after they occur. Here, we focus on two valuable ex ante modes of regulating algorithmic management that are currently in force: pre-emptive duties on employers under data protection and equality law and collective bargaining over the introduction and use of technology at work. We regard these as mutually reinforcing mechanisms that

⁴⁷ European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD), Annex III, para 4 <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

⁴⁸ EU AI Act, Ch 2 and art 16.

⁴⁹ From the labour sphere, see Valerio de Stefano, 'The EU Proposed Regulation on AI: A Threat to Labour Protection?' (*Regulating for Globalization Blog*, 2021) <<http://global-workplace-law-and-policy.kluwerlawonline.com/2021/04/16/the-eu-proposed-regulation-on-ai-a-threat-to-labour-protection/>>; and Aislinn Kelly-Lyth, 'Dispatch 39—European Union: The AI Act and Algorithmic Management' (2021) Comparative Labor Law & Policy Journal 1.

could be used to improve the position of workers who are subject to algorithmic management.

Comparable to the AI Act's self-assessment approach, there are existing pre-emptive duties on employers to conduct assessments of any proposed algorithmic management strategy in advance of its introduction. In the United Kingdom (UK), for example, there is a specific duty imposed on all public sector actors to have regard to equality considerations when making decisions,⁵⁰ and this should guide and constrain the implementation of algorithmic management in the public sector. Similarly, employers' duty to make reasonable adjustments for workers with a disability means they must take steps to ensure that the implementation of any workplace technology does not disadvantage them.⁵¹ Of more general relevance to these data-driven technologies is the duty, imposed by article 35 of the General Data Protection Regulation (GDPR), to conduct impact assessments where technologies pose a high risk to the rights of data subjects.

Algorithmic management of the kind outlined herein falls within the scope of the GDPR's listed situations that require a data protection impact assessment (DPIA).⁵² The assessment must include, amongst other content, an assessment of the risks to the rights and freedoms of data subjects and outline the measures envisaged to address the risks. A broad interpretation of the phrase 'rights and freedoms' is supported by the recommendations of the Article 29 Data Protection Working Party.⁵³ We argue that it should be taken as referring to the rights contained in the EU Charter of Fundamental Rights,⁵⁴ as well as those in the ECHR as discussed above. In a recent case relating to a Police authority's use of facial recognition software, the UK Court of Appeal held that, despite attempting to address the ECHR, art 8 implications of its high-risk processing, the DPIA 'failed properly to assess the risks to the rights and freedoms of data subjects and failed to address the measures envisaged to address the risks' as required by the GDPR and Data Protection Act.⁵⁵ In the context of algorithmic management tools, therefore, employers are required to undertake an assessment of the risk to a wide range of human rights and identify steps to minimise them. Failure to comply with the obligation to conduct a DPIA may result in a significant fine.

Soft law and self-regulatory measures such as impact assessments will not prevent all infringements of workers' fundamental rights by algorithmic management.⁵⁶ Nevertheless, these assessments do have the potential to raise awareness

⁵⁰ Equality Act 2010, s 149 (UK).

⁵¹ *ibid* ss 39(5), 20.

⁵² See GDPR, art 35(3)(a).

⁵³ Article 29 Data Protection Working Party, 'Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks' (WP 218, 30 May 2014) 4.

⁵⁴ Heleen Janssen, 'An Approach for a Fundamental Rights Impact Assessment to Automated Decision-Making' (2020) 10 International Data Privacy Law 76.

⁵⁵ *R (Bridges) v Chief Constable of South Wales Police* [2020] EWCA Civ 1058 [152].

⁵⁶ See the chapter by Alessandro Ortalda and Paul De Hert in this volume.

amongst employers of likely impacts upon human rights and to help avoid unintentional violations from taking place. Provided, of course, that they are taken seriously by employers and not regarded as a mere tickbox exercise. One significant shortcoming in respect of the current law's effectiveness is that the assessments conducted under the GDPR are not publicly accessible, making it difficult for workers, unions, and external organisations to audit them. This is problematic, as access to and scrutiny of impact assessments is key if they are to make a valuable contribution to the protection of human rights at work. This is therefore an area where further reform is needed, for example, through the introduction of independent external auditing requirements.⁵⁷ It is unfortunate that, while there is some provision for external certification of this kind in the EU's draft AI Act, it does not take the opportunity to extend certification to the use of algorithmic management of work.⁵⁸

If DPIAs were publicly accessible, they could form part of a virtuous circle with the second important mode of ex ante regulation identified here: collective negotiation and agreements between employers and trade unions. As Valerio de Stefano argues, in addition to setting out adequate standards for the treatment of the worker, regulation of technology in the workplace must be adaptable and adapted to the needs of specific workplaces.⁵⁹ Collective bargaining and agreements provide a flexible and context-specific means of shaping the algorithmic management of workers.⁶⁰ Trade unions or worker representatives are in a position to highlight the threats that workers face in their particular workplace, which are likely to include those set out above, as well as more context-specific threats to fair, decent, and safe working conditions. For example, union representatives may be able to resist an algorithm that sets an expected pace of work at 95 per cent of the quickest worker, on the grounds that such a high standard is inappropriate and impractical given that the workforce contains a range of physical abilities. In this way, collective agreements reached between unions and employers can complement more abstract national regulation by introducing a framework that is tailored to counter the distinctive risks that arise in a particular workplace or sector.

Indeed, collective negotiations and agreements relating to algorithmic management have already begun. In Spain, for instance, the government reached an agreement with social partners earlier this year on the rights of platform workers.⁶¹ It requires platforms to share information about how working conditions are determined by mathematical or algorithmic formulae with the legal representatives

⁵⁷ Ifeoma Ajunwa, 'An Auditing Imperative for Automated Hiring Systems' (2021) 34 Harvard Journal of Law & Technology 622.

⁵⁸ Kelly-Lyth (n 49) 9.

⁵⁹ Valerio De Stefano, '"Negotiating the Algorithm": Automation, Artificial Intelligence, and Labor Protection' (2019) 41 Comparative Labor Law & Policy Journal 15, 30.

⁶⁰ *ibid* 31.

⁶¹ Ane Aranguiz, 'Spain's Platform Workers Win Algorithm Transparency' (*Social Europe*, 18 March 2020) <<https://socialeurope.eu/spains-platform-workers-win-algorithm-transparency>>.

of workers. The legislation is a huge boost for the position of unions in relation to bargaining to improve the conditions and treatment of workers in the sector. In the UK, public sector workers in Wales are now covered by a set of principles on ‘Digitalisation at Work’ agreed between trade unions, public sector employers, and the Welsh government. Key principles emphasised in this agreement are the centrality of worker voice and consultation when introducing new technology in the workplace, that implementation must be managed in such a way so as not to negatively impact workers’ health or wellbeing, and that workers’ rights must be safeguarded in the design and implementation of new technology.⁶² Whilst these changes were realised due to government action, there have also been successful negotiations over algorithmic management practices in the private sector, with similar negotiations and agreements existing in the logistics and transport sectors.⁶³

These examples show that it is possible for social partners and collective negotiations to lead to specific regulations that strike a fair balance between the rights of workers and the desire of employers to reap the benefits of new technologies. Of course, for collective bargaining to be an effective means of regulating algorithmic management it must be facilitated by supportive legal frameworks and reinforced by the ability of workers to take industrial action in disputes over the use of workplace technology. For example, a UK union, Independent Workers of Great Britain, staged protests against the introduction of fingerprint scanning for clocking in or out. Their resistance was successful: the employer halted the policy’s implementation.⁶⁴ To realise the potential of collective bargaining for influencing the integration of technology and AI in the workplace, it is vital that legal mechanisms exist to enable workers to exercise voice over the use of algorithmic management and exert pressure on employers by going on strike. This indicates the importance of further research into how existing legal frameworks can be leveraged to allow workers to participate in decisions relating to algorithmic management and thereby ensure the use of these technologies is consistent with their human rights.⁶⁵

⁶² See Workforce Partnership Council, ‘Agreement Partnership and Managing Change’ (2021) <<https://gov.wales/sites/default/files/publications/2021-12/workforce-partnership-council-agreement-2021.pdf>>.

⁶³ See Unite the Union, ‘Draft New Technology Agreement’ (2017) <www.unitetheunion.org/media/1236/draft-new-technology-agreement-october-2016.pdf>; and Communication Workers Union, ‘Key Principles Framework Agreement’ (2018) <www.cwu.org/wp-content/uploads/2020/12/Joint-draft-KEY-PRINCIPLES-FRAMEWORK-AGREEMENT_18_12_20_Final.pdf>.

⁶⁴ Ben Chapman, ‘UCL Strike: Outsourced Workers to Walk Out in Protest Over “Bullying and Discrimination”’ *Independent* (6 November 2019).

⁶⁵ On this question see Philippa Collins and Joe Atkinson, ‘Worker voice and algorithmic management in post-Brexit Britain’ (2023) 29 *Transfer: European Review of Labour and Research* 37.

5 Conclusion

In this chapter, we have provided an overview of the current uses of AI and related technologies by employers for the purposes of algorithmic management and highlighted how these pose a fundamental threat to human rights in the workplace. The risk of employers infringing workers' human rights through exercises of their discretion and managerial prerogative does not recede and, indeed, becomes more acute, where decisions are fully or partially automated. A human rights framing of algorithmic management can also assist us in developing a detailed regulatory response to this phenomenon, including by placing clear limits upon the use of technologies in the workplace and protecting workers from invasive or unfair uses of surveillance or management by AI. Further research is required to investigate the particular duties and remedies that should be put in place to protect rights in this context.⁶⁶ But these measures must be complemented by pre-emptive consideration of context-specific risks and commitments to take steps to mitigate those risks: collective bargaining and impact assessments are important tools for employers, unions, and workers in this process. While these alone will not be sufficient to secure workers' human rights against the full-frontal threat posed by new technologies, they are nevertheless central elements of the broader package of measures needed to govern AI in the workplace.

⁶⁶ For a proposed comprehensive statute protecting human rights at work, see Collins (n 38).

Artificial Intelligence and the Right to Health

Enrique Santamaría Echeverría

1 Introduction

One of the many promises of artificial intelligence (AI) is the development of human health-related applications and technologies. AI interventions for health, in the same way as other AI applications in different domains, have a positive and an adverse impact on human rights in general, and on the right to health, specifically.

Although there is a lack of consensus on the definition of health,¹ several international and supranational instruments have pursued the legal crystallisation of a right to health. At the international level, article 12 of the International Covenant on Economic, Social and Cultural Rights (ICESCR)² recognises the right of everyone to the enjoyment of the highest available standard of physical and mental health. At the regional level, different conventions and treaties have recognised the right to health. It is the case of the African Charter on Human and Peoples' Rights (ACHPR)³ and the additional protocol to the American Convention on Human Rights (ACHR) in the area of economic, social, and cultural rights.⁴ At European level, despite the absence of a specific right to health in the European Convention on Human Rights⁵ (ECHR), the European Court of Human Rights (ECtHR) has repeatedly protected it in its case law.⁶ At EU level, the Charter of Fundamental

¹ The World Health Organization's definition of health is as follows: '[A] state of complete physical, mental and social well-being and not merely the absence of disease or infirmity'. See World Health Organization (WHO), Constitution of the World Health Organization (1946) 2, <<https://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf?ua=1>>.

² International Covenant on Economic, Social and Cultural Rights (adopted 16 December 1966, entered into force 3 January 1976) 2200A (XXI) (ICESCR).

³ African Charter on Human and Peoples' Rights (adopted 27 June 1981, entered into force 21 October 1986) (1982) 21 ILM 58 (ACHPR).

⁴ Additional Protocol to the American Convention on Human Rights in the Area of Economic, Social and Cultural Rights (Protocol of San Salvador) (entered into force 16 November 1999) OAS Treaty Series No 69 (1988) reprinted in Basic Documents Pertaining to Human Rights in the Inter-American System OEA/Ser L V/II.82 Doc 6 Rev 1 at 67 (1992).

⁵ Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR).

⁶ For an overview of health-related issues in the case law of the ECtHR, see Council of Europe, European Court of Human Rights, *Health-Related Issues in the Case-Law of the European Court of Human Rights* (2005).

Rights of the European Union⁷ (CFREU) protects the right to respect the physical and mental integrity of the person (article 3), and the right of access to preventive health care and the right to benefit from medical treatment (article 35).

As can be already inferred from these provisions, the right to health is a two-dimensional right. One dimension protects the individual right to health and access to healthcare. The other dimension encompasses the social dimension of the right (notably, public and population health). AI applications impact—sometimes simultaneously—both dimensions.

Moreover, the right to health is an inclusive right. It considers a wide range of factors which enable the full enjoyment of the right (the social determinants of health, for instance, water, a healthy environment, safe food),⁸ and contains freedoms (such as the freedom from non-consensual treatment) and entitlements (eg to participate in the decision affecting the person's health).⁹

Although some of the challenges posed by using AI in the realm of human health may not be unique when compared to the use of this technology in other fields, there are certain specificities that need to be addressed within the framework of the protection of the right to health and other related rights (such as non-discrimination).¹⁰ The general aim of this chapter is to explore these challenges and specificities. For this purpose, it maps the most relevant AI applications in the field of health so as to subsequently (section 2) identify the main opportunities and challenges (section 3) derived from these applications. This chapter finally addresses some of the contemporary policy and regulatory proposals with regards to AI in health and presents several paths forward (section 4).

2 Mapping AI Health-Related Applications

AI technologies have found application in at least the following, often intertwined, fields: diagnosis and disease identification, personalised treatment, mental health, digital phenotyping, public and populations health, translational research, clinical care and healthcare management, and language understanding. This section presents a brief overview of these applications and the multiple opportunities for improving health at the individual and collective level.

Training AI algorithms requires vast amounts of information from which they can learn. It is no surprise that the most advanced AI applications for diagnosis

⁷ Charter of Fundamental Rights of the European Union [2010] OJ C83/389.

⁸ See eg UN Committee on Economic, Social and Cultural Rights (CESCR), 'General Comment No 14: The Right to the Highest Attainable Standard of Health (Article 12 of the Covenant)' (2000), E/C.12/2000/4 11 August 2000.

⁹ The right to non-discrimination is an essential component of the right to health. See WHO and Office of the United Nations High Commissioner for Human Rights (OHCHR), 'The Right to Health. Fact Sheet No 31' (2008) 1, 3.

¹⁰ *ibid* 7.

and disease identification are focused on image intense medical specialities.¹¹ Radiology was the first speciality to generate large amounts of digital data.¹² These data have been used, for example, to design algorithms to identify different pathologies through X-ray analysis or to create diagnostic software to detect wrist fractures in adults.¹³ In oncology, machine learning (ML) has helped reducing the cases of benign surgeries by predicting whether a high-risk lesion identified by a biopsy could become cancer at surgery.¹⁴ In ophthalmology, IDx-DR—one of the earliest cases of the US Food and Drug Administration (FDA) approved ML software for clinical care—can detect diabetic retinopathy in adult patients diagnosed with diabetes.¹⁵ In dermatology, phone applications like SkinVision can be used for triage to determine whether a visit to the dermatologist is necessary.¹⁶ Recent developments in AI are also very promising in the field of digital pathology.¹⁷

AI technologies have also found fertile ground to flourish in the field of personalised treatment. Supervised learning systems can offer an array of diagnosis from which, taking into account the particularities of the patient and the likelihood of that particular person developing a disease, a physician can choose.¹⁸ AI could also help processing data from various sources to create a unique data and health profile of every individual patient.¹⁹ AI interventions could be particularly useful for precision medicine,²⁰ an approach to treatment in which different factors and data sources are considered: behavioural data (eg lifestyle), sociomarkers (eg environment), biomarkers (such as genes), and other patient data.²¹ A recent example of the use of AI in precision health is an algorithm designed to integrate whole genome sequencing with Electronic Health Records (EHR) to study abdominal aortic aneurism.²²

¹¹ Thomas M Maddox, John S Rumsfeld, and Philip RO Payne, 'Questions for Artificial Intelligence in Health Care' (2019) 321(1) *JAMA: Journal of the American Medical Association* 31, 31.

¹² Michael van Hartskamp and others, 'Artificial Intelligence in Clinical Health Care Applications: Viewpoint' (2019) 8(2) *Interactive Journal of Medical Research* 1, 3.

¹³ Nariman Noorbakhsh-Sabet and others, 'Artificial Intelligence Transforms the Future of Health Care' (2019) 132(7) *American Journal of Medicine* 795; and Sara Gerke, Timo Minssen, and Glenn Cohen, 'Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare' in Adam Bohr and Kaveh Memarzadeh (eds), *Artificial Intelligence in Healthcare* (Academic Press 2020) 295, 298.

¹⁴ Emmanuel Fombu, *Predictive Medicine: Artificial Intelligence and its Impact on Healthcare Business Strategy* (Business Expert Press 2020) 39.

¹⁵ Blake Murdoch, 'Privacy and Artificial Intelligence: Challenges for Protecting Health Information in a New Era' (2021) 22(5) *BMC Medical Ethics* 122, 122; Gerke, Minssen, and Cohen (n 13) 298.

¹⁶ Angel N Desai, 'Artificial intelligence: Promise, Pitfalls, and Perspective' (2020) 323(24) *JAMA: Journal of the American Medical Association* 2448, 2449.

¹⁷ Van Hartskamp and others (n 12) 3.

¹⁸ Fombu (n 14) 44.

¹⁹ *ibid* 119.

²⁰ Desai (n 16) 2448.

²¹ Arash Shaban-Nejad, Martin Michalowski, and David L Buckeridge, 'Health Intelligence: How Artificial Intelligence Transforms Population and Personalized Health' (2018) 1(2) *NPJ Digital Medicine* 53, 53; Ilona Kickbusch and others, 'The Lancet and Financial Times Commission on Governing Health Futures 2030: Growing Up in a Digital World' (2021) 398b(10312) *The Lancet* 1727, 1739.

²² Noorbakhsh-Sabet and others (n 13) 798.

In the field of mental health, AI has been used to create bots and virtual therapists. This is the case of Woebot, an AI chatbot that offers cognitive behavioural therapy for anxiety and depression,²³ or Ellie, a virtual therapist that can read body language.²⁴ AI has also been used to anticipate mental problems and to predict major depressive disorders by using image heatmap pattern recognition.²⁵

Digital phenotyping,²⁶ an emerging field focusing on using background and passive data collected by our phones and social media to infer and study health status, insights, and dynamics,²⁷ has also found application in mental health. These passive and background data include speed, keyboard accuracy, phone habits, how we hold our phone, app updates frequency, and battery recharge.²⁸

Some digital phenotyping studies have already focused on depression,²⁹ Parkinson's, Alzheimer's,³⁰ and schizophrenia.³¹ Facebook (Meta), for example, announced an algorithm to scan users' posts for signs of suicidal thoughts and alert a specialised team.³² Another company, Mindstrong Health, has developed an app to use digital phenotyping data to detect moods and memory changes.³³ Other studies have used publicly available YouTube videos of children with autism to train AI algorithms to detect autism from body movements like tremors and seizures.³⁴

²³ Fombu (n 14) 120.

²⁴ ibid 128.

²⁵ Noorbakhsh-Sabet and others (n 13) 97 with further references.

²⁶ On digital phenotyping, see Jukka-Pekka Onnela, 'Opportunities and Challenges in the Collection and Analysis of Digital Phenotyping Data' (2021) 46(1) *Neuropsychopharmacology* 45; Kit Huckvale, Venkatesh Svetha, and Helen Christensen, 'Toward Clinical Digital Phenotyping: A Timely Opportunity to Consider Purpose, Quality, and Safety' (2019) 2(1) *NPJ Digital Medicine* 1.

²⁷ Sidney Fussell, 'YouTube Videos Are a Gold Mine for Health Researchers' (*The Atlantic*, 1 September 2019) <www.theatlantic.com/technology/archive/2019/09/breakthrough-autism-research-uses-social-media-videos/597646/>; Miron Dechansky, 'Digital Phenotyping: Turning our Smartphones Inward' (*Medium*, 11 April 2019) <<https://medium.com/leolab/digital-phenotyping-turning-our-smartphones-inward-141a75b2f2a3>>.

²⁸ Dechansky (n 27).

²⁹ Fabian Wahl and others, 'Mobile Sensing and Support for People with Depression: A Pilot Trial in the Wild' (2016) 4(3) *JMIR mHealth and uHealth* 5960.

³⁰ Umathi Reddy, 'Clues to Parkinson's and Alzheimer's From How You Use Your Computer' (*The Wall Street Journal*, 29 May 2028) <www.wsj.com/articles/clues-to-parkinsons-disease-from-how-you-use-your-computer-1527600547>.

³¹ Ian Barnett and others, 'Relapse Prediction in Schizophrenia through Digital Phenotyping: A Pilot Study' (2018) 43 *Neuropsychopharmacology* 1660.

³² Benjamin Goggin, 'Inside Facebook's Suicide Algorithm: Here's How the Company Uses Artificial Intelligence to Predict Your Mental State From Your Posts' (*Business Insider Nederland*, 6 January 2019) <www.businessinsider.nl/facebook-is-using-ai-to-try-to-predict-if-youre-suicidal-2018-12?international=true&r=US>.

³³ Poornima Venkatesan, 'Digital Phenotyping: A Revolution or a Privacy Breach' (*MedCity News*, 13 January 2019) <<https://medcitynews.com/2019/01/digital-phenotyping-a-revolution-or-a-privacy-breach/>>.

³⁴ For an example of these type of studies, see Andrew Cook and others, 'Towards Automatic Screening of Typical and Atypical Behaviors in Children with Autism' (2019) IEEE International Conference on Data Science and Advanced Analytics 504; Fussell (n 27). However, the repurposing of public available data raises difficult questions regarding consent from minors and adults.

Although the focus of digital phenotyping has been predominantly on individual's health, it has also found application in population health, public health, and public health surveillance.³⁵ It could be used, for example, for risk assessment and predict which kind of medical services and resources are required in a certain area³⁶ or for the purposes of digital epidemiology.³⁷ An example of the success of this type of applications is HealthMap, a real-time public health surveillance system, which, through the mining of internet sources, detected the 2014 Ebola outbreak before public authorities officially announced it.³⁸

Beyond digital phenotyping applications, AI has also been used in the area of public health³⁹ to predict epidemic outbreaks,⁴⁰ for population health management,⁴¹ and for precision public health.⁴² Some examples of epidemic prediction include detection of filoviruses outbreaks in Uganda,⁴³ cholera outbreaks in South Africa, and the calculation of the morbidity rate of dengue haemorrhagic fever in Thailand.⁴⁴ For population health management AI could be used to use individual data and sociomarkers to increase disease surveillance (also through the use of m-health⁴⁵), and prediction, implementation, and evaluation of population health interventions.⁴⁶

AI has also significant applications in translational research. Due to the high costs of drug development and the low success rate of clinical trials,⁴⁷ AI can be used for drug discovery and the screening of drug compounds.⁴⁸ Machine learning also finds applications in 'in silico clinical trials', being clinical trials using virtual patients. In silico trials and generative models whereby synthetic patient data is generated without any connection to real individuals,⁴⁹ allow researchers

³⁵ Shaban-Nejad, Michalowski, and Buckeridge (n 21) 52.

³⁶ Dechansky (n 27).

³⁷ Nicole Laskowski, 'Digital Epidemiology Takes the Flu with the Help From You' (*TechTarget*, 30 January 2018) <www.techtarget.com/searchcio/blog/TotalCIO/Digital-epidemiology-takes-on-flu-with-help-from-you>.

³⁸ Venkatesan (n 33).

³⁹ See Nina Schwalbe and Brian Wahl, 'Artificial Intelligence and the Future of Global Health' (2020) 395 *The Lancet* 1579; Tristan Panch and others, 'Artificial Intelligence: Opportunities and Risks for Public Health' (2019) 1(1) *Lancet Digit Health* 13, 13.

⁴⁰ Fombu (n 14) 45.

⁴¹ ibid 124.

⁴² Kickbusch and others (n 21) 1739.

⁴³ UNHRC, 'Annual Report of the United Nations High Commissioner for Human Rights and Reports of the Office of the High Commissioner and the Secretary-General, Question of the Realization of Economic, Social and Cultural Rights in All Countries: The Role of New Technologies for the Realization of Economic, Social and Cultural Rights' UN Doc A/HRC/43/29 (2020), 3 with further references.

⁴⁴ Noorbakhsh-Sabet and others (n 13) 798 with further references.

⁴⁵ Caroline Free and others, 'The Effectiveness of M-health Technologies for Improving Health and Health Services: A Systematic Review Protocol' (2010) 3.1 *BMC Research Notes* 1.

⁴⁶ Shaban-Nejad, Michalowski, and Buckeridge (n 21) 52.

⁴⁷ Noorbakhsh-Sabet and others (n 13) 797.

⁴⁸ Fombu (n 14) 45.

⁴⁹ Murdoch (n 15) 125. The use of generative and synthetic data would, for obvious reasons, avoid some of the privacy related issues with the use of health and other personal and sensitive data.

to partially replace humans and animals, and are particularly helpful with disease with limited data availability, for instance, trials for orphan diseases.⁵⁰

AI technologies are, moreover, relevant in clinical care and healthcare management. A paradigmatic case is triage.⁵¹ AI-powered apps could help determining how urgent a case is: Ada, for example, provides guidance about the level of a certain emergency; and Corti, a ML software program, helps emergency dispatchers make decisions by analysing voice, symptoms, breath patterns, and other metadata.⁵² AI can also be used to save time in healthcare⁵³ through operational efficiency and performance,⁵⁴ automated planning and scheduling,⁵⁵ and automated clinical decisions systems.⁵⁶ Moreover, by relying on AI for the integration of data and knowledge into clinical workflow, physicians can devote more time to their patients and less on the administrative side of the work.⁵⁷

Finally, AI intervenes in the field of health through language understanding applications.⁵⁸ By means of natural language processing, algorithms allow machines to identify key words or sentences in unstructured written texts (eg clinicians' notes)⁵⁹ and understand their meaning.⁶⁰ A famous example of this type of health-related application is Amazon Comprehend Medical.⁶¹ According to their website, this ML application can 'extract medical information from unstructured medical text like doctors' notes, clinical trial reports, or radiology reports'.⁶² AI for language processing may also be applied to biomedical text mining and sentiment analysis in internet-derived data.

Mapping different AI applications for health serves as a methodological tool to identify the human rights challenges posed by these technologies. In the same way as the right to health can be explained by appealing to its two-dimensional nature, AI interventions may impact both dimensions of this right. From this mapping, a clear promise can be distilled: the improvement of human health at the individual and collective level. Furthermore, AI for health could play a role in addressing global health inequalities by advancing and transforming (public) health

⁵⁰ Another advance of these types of clinical trials is that they overcome privacy issues raised by the use of personal (health) data.

⁵¹ Desai (n 16) 2449.

⁵² Gerke, Minssen, and Cohen (n 13) 300 with further reference.

⁵³ Fombu (n 14) 116.

⁵⁴ *ibid* 124.

⁵⁵ Brian Wahl and others, 'Artificial Intelligence (AI) and Global Health: How Can AI Contribute to Health in Resource-Poor Settings?' (2018) 3(4) *BMJ Global Health* 1, 2.

⁵⁶ The combination of human and AI has been termed augmented intelligence. It has been argued that augmented intelligence works better than human intelligence or AI alone. See Noorbakhsh-Sabet and others (n 13) 798.

⁵⁷ Maddox, Rumsfeld, and Payne (n 11) 32.

⁵⁸ Marcel Salathé, Thomas Wiegand, and Markus Wenzel, 'Focus Group on Artificial Intelligence for Health' (2018) World Health Organization 3.

⁵⁹ Maddox, Rumsfeld, and Payne (n 11) 31.

⁶⁰ Wahl and others (n 55) 2; Desai (n 16) 2448.

⁶¹ See <<https://aws.amazon.com/comprehend/medical/>>.

⁶² *ibid*.

in low- and middle-income countries (LMIC), remote areas,⁶³ or resource-poor settings.⁶⁴ AI-powered low-cost tools running on phones would help allocating scarce resources (eg triage for skin cancer), support clinical decision-making, and improve public health by determining the relations of causality in epidemics.⁶⁵ Optolexia, for example, is a ML algorithm that can detect dyslexia in small children through the use of a laptop and an eye tracker.⁶⁶ In resource-poor settings with strong internet and high laptop and mobile penetration rates, AI can be used to tackle problems derived from the lack of physicians or through the use of expert systems to predict, model, and slow the spread of diseases in epidemic situations.⁶⁷ AI may also be helpful in improving the efficiency of immunisation initiatives, supply chain, and referral services.⁶⁸

Other human rights opportunities in health are alleviating health disparities with AI. AI could help remedying the traditional underrepresentation of women and minorities in health research and innovation,⁶⁹ and it could work as a support tool to address possible physician biases.⁷⁰ Finally, AI in clinical, translational, and educational settings can help expanding knowledge and research,⁷¹ and improve medical education and training.⁷²

However, besides these major beneficial consequences for health-related human rights, AI technologies pose also challenges and threats to the enjoyment of the right to health and other human rights. Section 3 explores these challenges.

3 Challenges

Some of the problems encountered in the intervention of AI technologies in health are common to the ones encountered in the deployment of AI technologies in other fields. However, there are certain specificities of health AI applications and the right to health that need to be particularly addressed. In the same way as the right to health encompasses an individual and a social dimension, the challenges may impact more one or another dimension of the right (including social determinants

⁶³ Ahmed Hosny and Hugo JWJ Aerts, ‘Artificial Intelligence for Global Health’ (2019) 366(6468) Science 955.

⁶⁴ Wahl and others (n 55) 2.

⁶⁵ Hosny and Aerts (n 63) 955.

⁶⁶ Fombu (n 14) 125.

⁶⁷ For example, a ML tool to identify weather and land-use patterns associated with dengue fever transmission in Manila. See Wahl and others (n 55) 4; Venkatesan (n 33).

⁶⁸ Wahl and others (n 55) 4.

⁶⁹ Irene Y Chen, Shalmali Joshi, and Marzyeh Ghassemi, ‘Treating Health Disparities with Artificial Intelligence’ (2020) 26(1) Nature Medicine 16 with further references and examples.

⁷⁰ Ravi B Parikh, Stephanie Teeple, and Amal S Navathe, ‘Addressing Bias in Artificial Intelligence in Health Care’ (2019) 322(24) JAMA: Journal of the American Medical Association 2377, 2378.

⁷¹ UNHRC (n 43) 5.

⁷² Michael J Rigby, ‘Ethical Dimensions of Using Artificial Intelligence in Health Care’ (2019) 21(2) AMA Journal of Ethics 121.

of health, freedoms, and entitlements). These challenges can be clustered in four different but related headings:

- (i) Inaccuracy;
- (ii) Unexplainably and opacity;
- (iii) Threats to privacy which indirectly impact the right to health;
- (iv) Other issues.

3.1 Inaccuracy of AI Applications

AI algorithms and health data may be biased and prone to errors. Data may be biased for several reasons. Physicians and nurses may prefer to record information consistent with their previous knowledge, experiences, or prejudices,⁷³ neglecting other types of data.⁷⁴ Even in the absence of prejudices, data may be biased for the place where it was collected or where the model was trained.⁷⁵ The collection of biased data can also occur by oversampling sicker populations (eg by only taking into account clinical data) or healthier populations (by the use of data extracted exclusively from wellness devices and applications).⁷⁶ Contextual bias can also occur despite the use of high-quality data:⁷⁷ a given recommendation can work perfectly from a purely medical perspective, but is likely not to work or be deadly in a low-resource environment or a low- or middle-income country because of lack of resources to cope with the medical implication of AI recommendations. Contextual biases are particularly hard to detect in opaque AI.

The use of biased data in AI applications may have various nefarious human rights consequences: it may perpetuate existing inequalities related to ethnic, gender, or socio-economic factors⁷⁸ and maintain the underrepresentation of minority groups in data sets.⁷⁹ Moreover, AI models developed in high income countries⁸⁰ and developed for prevalent diseases may not be of any utility in other types of settings or for other type of diseases.⁸¹

These negative effects derived from the inaccuracy of AI have a clear translation in terms of the right to health: certain groups within a society or entire countries

⁷³ I Glenn Cohen and others, ‘The European Artificial Intelligence Strategy: Implications and Challenges for Digital Health’ (2020) 2(7) *The Lancet Digital Health* 367, 377.

⁷⁴ Marzyeh Ghassemi and others, ‘Practical Guidance on Artificial Intelligence for Health-Care Data’ (2019) 1(4) *The Lancet Digital Health* 157, 157; see also Parikh, Teeple, and Navathe (n 70) 2377.

⁷⁵ Noorbakhsh-Sabet and others (n 13) 799.

⁷⁶ Maddox, Rumsfeld, and Payne (n 11) 31.

⁷⁷ Timo Minssen and others, ‘Regulatory Responses to Medical Machine Learning’ (2020) 7(1) *Journal of Law and the Biosciences* 1, 17.

⁷⁸ Wahl and others (n 55) 5.

⁷⁹ Panch and others (n 39) 13.

⁸⁰ Hosny and Aerts (n 63) 955.

⁸¹ Desai (n 16) 2448.

and populations may receive substandard access to AI applications and health-care, if any. Moreover, as it has been pointed out before, since the right to non-discrimination constitutes one of the essential components of the right to health,⁸² its violation may directly impact the right to health.

In any case, the question on how to deal with the potential discriminatory effects of AI algorithms remains. Some scholars have asserted that the right to non-discrimination may not be of much help in solving AI problems in the context of health since including, for example, race as a variable may produce more racially equitable health outcomes.⁸³

3.2 Unexplainability and Opacity

Some AI technologies provide insights via unobservable methods.⁸⁴ In fact, learning algorithms methods can be partially or completely opaque to human observers.⁸⁵ This is particularly true for deep learning models which are especially hard to interpret and explain.⁸⁶ This phenomenon of opaqueness has been labelled the black box problem.⁸⁷

The black box poses challenges for the processes of informed consent and the right to an explanation.⁸⁸ Both these rights are closely linked with human dignity and the right to health.⁸⁹ In fact, the right to an informed consent constitutes one of the entitlements derived from the right to health.⁹⁰ How can a physician or a researcher obtain informed consent for the use of AI technologies for health when she is not capable of explaining how the technology works and produces outcomes?

Some of these problems also reflect on a broader question for society at large, namely, what is the right amount of opacity we are willing to accept in exchange for accuracy. In other words, is it desirable to find explainability sacrificing innovation?⁹¹ At this point, the individual and collective dimensions of the right to health collide. One may argue that more innovation—sacrificing explainability—may lead to more and better technologies for health. This, in its turn, would benefit the social dimension of right to health. However, less explainability may affect the rights of individuals in the context of healthcare, including diagnosis and treatment.

⁸² WHO (n 9) 3.

⁸³ Cohen and others (n 73) 376.

⁸⁴ Maddox, Rumsfeld, and Payne (n 11) 32.

⁸⁵ Murdoch (n 15) 123.

⁸⁶ Salathé, Wiegand, and Wenzel (n 58) 3.

⁸⁷ Maddox, Rumsfeld, and Payne (n 11) 33.

⁸⁸ Gerke, Minssen, and Cohen (n 13) 301.

⁸⁹ European Union Agency for Fundamental Rights, ‘Getting the Future Right: Artificial Intelligence and Fundamental Rights’ (2020), 60.

⁹⁰ WHO (n 9) 3.

⁹¹ Cohen and others (n 73) 376.

A somehow related problem pertaining to the explainability and safety of AI algorithms for health is the so-called update problem. Certain medical products (devices and drugs) are required to undergo reviewing and approval by the appropriate authorities (such as the—US—FDA or—European—EMA) before they are marketed. This presents certain difficulties when it comes to ML algorithms to be used in the health domain. Does an updated AI need a new revision and approval or is the initial one sufficient?⁹² One of the proposed solutions has been to lock the algorithms in such a way that they do not change over time and do not use new data to change their own performance.⁹³ Once again, the challenges for the right to health can be framed in terms of trade-offs. Locking the algorithm would impede the exploitation of the full potential of AI learning algorithms, which would as well limit the potential benefits for the right to health.⁹⁴ However, unlocked algorithms may also pose significant risks to the safety⁹⁵ and health of patients and users of AI applications.⁹⁶

Regarding these safety and transparency⁹⁷ related problems, it is paramount to know how the models are being developed, and which (high-quality and robust) data sets are being used. This can in its turn (partially) address the well known problems of lack of public trust in AI for health.

3.3 Threats to Privacy Which Indirectly Impact the Right to Health

In addition to the problems that inaccuracy and opacity create for the right to health, AI technologies for health raise questions in relation to the protection of other human rights, namely, privacy and data protection.

It is well known that AI technologies need large amounts of data to be trained and operate. In the case of AI for health, these data are often sensitive data.⁹⁸ Perhaps even more significantly, data initially considered as not personal, through aggregation and AI analysis can reveal health traits of an individual. In fact, since it

⁹² Boris Babic and others, 'Algorithms on Regulatory Lockdown in Medicine' (2019) 366(6470) Science 1202, 1202.

⁹³ For an explanation of what an unlocked algorithm is, see Minssen and others (n 77) 1.

⁹⁴ Cohen and others (n 73) 377.

⁹⁵ On a similar note, a well known example of the safety problems of AI in health is IBM Watson for oncology, an AI system developed to help physicians with cancer diagnostics which was the object of strong criticism for providing unsafe recommendations for cancer treatment. See Gerke, Minssen, and Cohen (n 13) 302. For a brief overview of the problems with Watson and the unfulfilled promises of AI in the medial realm, see Steve Lohr, 'What Ever Happened to IBM's Watson' (*The New York Times*, 16 July 2021) <www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html>.

⁹⁶ According to recent data, the FDA has only approved medical AI devices with locked algorithms. See Minssen and others (n 77) with further references.

⁹⁷ Transparency may find limits by the existence of intellectual property rights and trade secrets.

⁹⁸ Noorbakhsh-Sabet and others (n 13) 799; Rigby (n 72) 122.

is possible to derive and predict health issues from social media data or the digital phenotype, the very notion of health data as a separate legal category has been challenged.⁹⁹ Moreover, the efficacy of traditional technical measures for the protection of privacy and identity (eg anonymisation, codification, pseudonymisation) is cracking by the use of computation strategies which, through the use of big data and aggregation, can easily re-identify individuals in health databases.¹⁰⁰

In addition, because of AI's 'lust' for data, novel questions arise regarding the secondary uses of health data. Depending on the purposes for which personal data are collected, they may have two types of uses: primary and secondary. The first are the uses for which the data were initially collected (eg in the context of clinical care). Secondary uses are those derived from further processing for purposes other than the initial ones. Secondary uses may vary. Particularly relevant are the processing of personal data for the purposes of public health protection, including the planning, management, administration, and improvement of health systems or the prevention and control of diseases, and the use of data in scientific, historical, or statistical research by public or private entities.

The distinction between primary and secondary uses of data is useful to determine the lawfulness of the processing of personal data. A general data protection principle establishes that the collection of data must answer to specific and well-defined purposes and that the further processing of the collected data must not be incompatible with the initial purposes for which data was collected. One of the rationales behind this principle is that, to exercise her autonomy and privacy rights, the data subject must know precisely, and in advance, what she is consenting to.¹⁰¹

The secondary use of data in the areas of public health and research reflects the tensions between public and private interest around the use of health data in AI: on the one hand, the interests of the data subject whose privacy may be violated if the data are used without her consent; on the other hand, there are the interests of society to advance research and protect and promote health. In principle, when there is consent for the further processing of data, there would not be a problem. The legal questions are more evident in the cases in which the further processing of the data occurs without the consent of the data subject. For which purposes would it be desirable to allow the secondary use of health data? An interesting, albeit not always clear solution has been to allow secondary uses of data for reasons

⁹⁹ Some scholars have proposed that all personal data should be treated as health data. See Christophe Olivier Schneble, Bernice Simone Elger and David Martin Shaw, 'Google's Project Nightingale Highlights the Necessity of Data Science Ethics Review' (2020) 12(3) EMBO Molecular Medicine 1, with further references.

¹⁰⁰ Murdoch (n 15) 124. The problems of re-identification were already spotted for the use of genomic data. Since genomic data is unique to every individual, even in the cases when genomic data was anonymised, it is possible, by crossing a couple of databases, to trace back the identity of the person.

¹⁰¹ At the EU level, this principle (purpose limitation) is enshrined in art 5(1)(b) of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (GDPR).

of public interest in the area of public health or for scientific research purposes.¹⁰² In this regard, a few questions arise: What constitutes research? And what does public interest entail? Does the industry carry out research in the public interest or must it be conducted by an academic institution or a healthcare entity? If so, under which circumstances? According to some opinions,¹⁰³ the industry (pharmaceutical, biotechnology, technology companies, medical technology industries, and insurance providers) can conduct research in the public interest. In this regard, the European Data Protection Supervisor (EDPS) has warned about the difficulties of distinguishing between genuine research and other types of research, which serve private interests.¹⁰⁴

Some of the risks to privacy and data protection are exemplified by the following cases involving big technological companies.

In 2016, DeepMind, a technology company in the UK, partnered with the Royal Free London NHS Foundation to develop a detection tool for acute kidney injury.¹⁰⁵ However, there were allegations of DeepMind acquiring the records of 1.6 million patients without an appropriate legal basis (including consent) and it was, in fact, found in breach of the UK Data Protection Act 1998¹⁰⁶ by the Information Commission.¹⁰⁷ DeepMind was subsequently bought by Google. Google (DeepMind) is now facing a class action for the unauthorised use of NHS medical records.¹⁰⁸

For project Nightingale, Google entered into a partnership with Ascension, one of the largest healthcare providers in the United States (US).¹⁰⁹ Through this partnership, Google received the medical data of around 50 million of Ascension's customers. The transferred data was not anonymised, and patients and doctors were not notified or asked for consent.¹¹⁰

In another case involving Google, the tech giant entered into a partnership with the University of Chicago to develop ML tools with medical application. For these purposes, the University of Chicago transferred hundreds of de-identified medical

¹⁰² This is the stance of art 9(2)(i)–(j) of the GDPR.

¹⁰³ Johan Hansen and others, *Assessment of the EU Member States' Rules on Health Data in the Light of GDPR* (Publications Office of the EU 2021) 57.

¹⁰⁴ European Data Protection Supervisor, 'A Preliminary Opinion on Data Protection and Scientific Research' (2020) 1, 3.

¹⁰⁵ Murdoch (n 15) 123; Gerke, Minssen, and Cohen (n 13) 305.

¹⁰⁶ Which implemented the old EU Data Protection Directive 95/46/EC now replaced by the GDPR.

¹⁰⁷ Gerke, Minssen, and Cohen (n 13) 305 with further references.

¹⁰⁸ Tammy Lovell, 'Google and DeepMind Face Legal Claims for Unauthorized Use of NHS Medical Records' (*Healthcare IT News*, 17 May 2022) <<https://www.healthcareitnews.com/news/emea/google-and-deepmind-face-legal-claim-unauthorised-use-nhs-medical-records>>; Natasha Lomas, 'Google Faces Fresh Class Action-Style Suit in UK over DeepMind NHS Patient Data Scandal' (*TechCrunch*, 16 May 2022) <<https://techcrunch.com/2022/05/16/google-deepmind-nhs-misuse-of-private-data-lawsuit/>>.

¹⁰⁹ Rob Copeland, 'Google's "Project Nightingale" Gathers Personal Health Data on Millions of Americans' *The Wall Street Journal* (11 November 2022).

¹¹⁰ Schneble, Elger, and Shaw (n 99) 1.

records to Google. In *Dinerstein v Google*,¹¹¹ Mr Dinerstein, the plaintiff and one of the University's patients, alleged, among other claims, breach of contract against Google and the University of Chicago for violation of the US Health Insurance Portability and Accountability Act (HIPPA). Although a motion to dismiss was granted in favour of the defendants, several voices have raised concerns about the lack of protection for health data under US law.¹¹²

Independently of the major privacy and data protection issues raised by the misuse of health data in AI applications, these examples also signal another big challenge related to AI in health, namely, the concentration of technological innovation and knowledge in big tech, with every major digital company investing in AI technologies for health.¹¹³

The growing accumulation of power and health data in the hands of big technological companies—via a relaxed understanding of the requirements for the lawful secondary use of data—may have unsought consequences for the social dimension of the right to health: it is likely to limit beneficial innovation for health systems and the public domain.

Besides the aforementioned normative problems, AI applications for health also pose specific technical and practical challenges regarding data. AI for health requires three famous Vs: volume, velocity, and variety.¹¹⁴ However, since the use of health data is regulated by data protection laws, access to training data is difficult, limiting the predictive capability of models.¹¹⁵ Moreover, for the development of high-quality AI models, it is necessary to improve the quality of data sources: most clinical data is incomplete,¹¹⁶ of bad quality, or hard to interpret for AI systems.¹¹⁷ In terms of data curation, problems of interoperability between different data sources and health institutions remain. It is still very difficult to use data across different health systems, with obvious consequences for the volume and variety variables.¹¹⁸ Until the aforementioned technical challenges are solved, the possibility

¹¹¹ *Dinerstein v Google* 484 F.Supp. 3d 561 (United States District Court, ND Illinois, Eastern Division, 2020).

¹¹² Jenna Becker, 'Insufficient Protections for Health Data Privacy: Lessons from Dinerstein v. Google' (*Bill of Health*, 28 September 2020) <<https://blog.petrieflom.law.harvard.edu/2020/09/28/dinerstein-google-health-data-privacy/>>.

¹¹³ Ezekiel J Manuel and Robert M Wachter, 'Artificial Intelligence in Health Care' (2019) 321(23) *JAMA: Journal of the American Medical Association* 2281, 2281. For a revealing text on the influence of BigTech in the medical sector, see the following interview (in Dutch) with philosopher Tamar Sharon: Marjolein van Trigt, 'Grote Techbedrijven infiltreren in de Medische Sector en dat is een Kwalijke Zaak' (*DeMorgen*, 6 July 2022) <www.demorgen.be/beter-leven/grote-techbedrijven-infiltreren-in-de-medische-sector-en-dat-is-een-kwalijke-zaak~b38ee49d/?utm_campaign=shared_earned&utm_medium=social&utm_source=whatsapp>.

¹¹⁴ Noorbakhsh-Sabet and others (n 13) 798.

¹¹⁵ Salathé, Wiegand, and Wenzel (n 58) 3 with further references.

¹¹⁶ Ghassemi and others (n 74) 158.

¹¹⁷ Van Hartskamp and others (n 12) 4.

¹¹⁸ Tristan Panch, Peter Szolovits, and Rifat Atun, 'Artificial Intelligence, Machine Learning and Health Systems' (2018) 8(2) *Journal of Global Health* 1, 5.

of moving from AI applications in relatively narrow domains to wider purpose multimodal AI systems seems distant.¹¹⁹

3.4 Other Issues

Some other threats to the right to health arise, not any more from AI technology itself, but from voluntary or involuntary human actions, including actions threatening cybersecurity, and human behaviour to overcompensate or overadjust to AI outputs (eg AI recommendations).

The cybersecurity concerns transcend the leak or hack of sensitive data. Hackers can now target not only institutions, but also body-internal and body-melded devices connected to the internet, putting at risk individual's health and physical integrity.¹²⁰

Finally, AI may have pervasive consequences when deployed for hospital management¹²¹ or clinical workflow. Doctors may be prone to overcompensating—which can also cause an increase in the cognitive burden of clinical teams¹²²—or to over-adjustment and automation complacency¹²³ to the detriment of patients' care and health.¹²⁴

4 Policy and Regulatory Proposals

AI applications create a constellation of opportunities for the improvement of human health, as well as various multilayered challenges negatively impacting human rights. It is therefore hard to find a one-size-fits-all solution for the problems described in section 3. However, there seems to be a consensus on the idea that the unique characteristics of AI technologies introduce unique regulatory challenges, and demand unique solutions¹²⁵ associated with a high level of oversight and regulations.¹²⁶

¹¹⁹ Van Hartskamp and others (n 12) 3.

¹²⁰ On the distinction between body-melded and body-internal devices, see Andrea M Matwyshyn, 'The Internet of Bodies' (2019) 61 William & Mary Law Review 77. This classification seems to follow very closely another classification proposed for the use of AI in the wellness field. Thus, it has been proposed to classify AI devices into intangible AI, tangible AI, and fused or embedded AI. Examples of the first type are applications for psychological therapy and the so-called smart hotel rooms. Examples of the second type are 'smart' sex dolls with the ability to adapt to the emotional and sexual tastes of their owner. Finally, examples of the third type are devices (invasive or non-invasive) that allow for Brain-Computer Interfaces. On this, see Lydia Kostopoulos, 'The Emerging Artificial Intelligence Wellness Landscape: Benefits and Potential Areas of Ethical Concern' (2018) 55(1) California Western Law Review 235.

¹²¹ Fombu (n 14) 41.

¹²² Maddox, Rumsfeld, and Payne (n 11) 32.

¹²³ Parikh, Teeple, and Navathe (n 70) 2377.

¹²⁴ European Union Agency for Fundamental Rights (n 89) 64.

¹²⁵ Murdoch (n 15) 122.

¹²⁶ Maddox, Rumsfeld, and Payne (n 11) 32.

Perhaps the most stimulating scholarly proposal regarding the use of AI and ML-based software as medical devices¹²⁷ has been advanced by Gerke, Babic, Evgeniou, and Cohen.¹²⁸ To solve the problems already described in this chapter, these scholars have signalled the need for a system view to regulate AI medical devices, instead of a product-centred regulation. Taking inspiration from the functions of the United Kingdom's specialised Human Fertilisation and Embryology Authority (HFEA) which licenses individual clinics for particular reproductive technologies, these scholars suggest regulators of AI medical devices should take a similar, but more challenging approach.¹²⁹ Regulators would need to take into account more aspects of health delivery and various technologies when authorising AI medical devices. Additionally, according to their proposal, 'such an approach would also raise difficult questions about how far upstream regulator would need to go, for example, in validating "golden data sets" as ground truth comparators'.¹³⁰ A systemic approach to AI medical devices would also imply constant analysis of the interaction and outcomes between humans (physicians and nurses) and AI systems and a continuous risk-monitoring approach¹³¹ to identify and manage risks associated with the use of AI in health.¹³² Although more demanding from the point of view of the regulators, this proposal would have beneficial consequences for both dimensions of the right to health. A strengthened AI oversight may reduce the materialisation of threats to individual's health, increasing at the same time public trust in AI technologies. At the social level, the validation of golden data sets and ground truths would also reinforce the protection of health-related human rights, such as the right to privacy and data protection.

In addition to this overarching proposal for the regulation of AI medical devices, other solutions have been advanced to more specific problems relating to AI interventions in health.

¹²⁷ Medical ML algorithms can be classified as software in a medical device or software as medical device. See Minssen and others (n 77) 5.

¹²⁸ Sara Gerke and others, 'The Need for a System View to Regulate Artificial Intelligence/Machine Learning-Based Software as Medical Device' (2020) 3(1) NPJ Digital Medicine 1. Two other comprehensive proposals on ethics and governance of AI for health are: World Health Organization, 'Ethics and Governance of Artificial Intelligence for Health: WHO Guidance' (2021); and ITU-T Focus Group on AI for Health (FG-AI4H), 'Ethics and Governance of Artificial Intelligence for Health (Version 1)' (2022).

¹²⁹ For clarity purposes I transcribe here the authors' description of the HFEA activities for authorisation: 'One could, perhaps, find some loose analogies in the way the United Kingdom's specialized Human Fertilisation and Embryology Authority (HFEA) licenses individual clinics for particular reproductive technology uses such as maternal spindle transfer (MST) or pronuclear transfer (PNT), two mitochondrial replacement techniques to prevent the transmission of serious mitochondrial disease from a mother to her infant. To act lawfully, clinics need to get a license from the HFEA to carry out one or both of these techniques—they need to show the capability to perform MST and/or PNT. In addition, they also must receive another approval from the HFEA when using one of these techniques for a particular patient.' See Gerke and others (n 128) 3.

¹³⁰ *ibid* 3.

¹³¹ Babic and others (n 92) 1204.

¹³² *ibid* 1202; Cohen and others (n 73) 377; Minssen and others (n 77) 5.

Regarding the legal—but also ethical and technical—issues associated with the use of health data, the necessity has been pointed out of more harmonised health infrastructures.¹³³ This would enable the sharing of health data between different stakeholders for the purposes of research and innovation and would also help tackle some of the current challenges regarding the variety and volume of data for AI applications in health. One example of a step forward in this direction is the EU Commission proposal for a Regulation on the European Health Data Space (EHDS),¹³⁴ already envisaged by the European Strategy for Data.¹³⁵ According to this strategy, the ‘EHDS will also promote better exchange and access to different types of electronic health data, including electronic health records, genomics data, patient registries etc’.

Although the EHDS proposal and the new Data Governance Act¹³⁶ seem well-intentioned, they have been the object of criticism for abandoning the well known notion of data solidarity and replacing it for the vaguer one of data altruism. The idea of data solidarity refers to the need of harnessing for the common good the potential economic, cultural, and social benefits derived from the sharing of (health) data.¹³⁷ While the notion of data solidarity evokes the paradigm of reciprocity,¹³⁸ the concept of data altruism exploits the idea of individuals selflessly sharing their data. Both the idea of data ownership and data altruism play well with the enormous power big technological companies have already accumulated in the field of digital technologies—including AI—in the realm of health. As has been pointed out by the editorial of the Lancet Commission on governing health futures, data solidarity ‘(r)ather than being regarded as something to be owned and hoarded, it emphasises the social and relational nature of health data’.¹³⁹

It would be desirable from a policy perspective to come back to the notion of data solidarity as an underpinning for the use of health data in and for AI technologies. Unlike data altruism, data solidarity stresses the relevance of returning to society some of the benefits to the right to health harnessed by the use of health data for the development of AI (and other) technologies.

¹³³ Cohen and others (n 73) 378.

¹³⁴ See European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space’ COM(2022) 197 final (3 May 2022).

¹³⁵ See European Commission, ‘A European Strategy for Data: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions’ COM(2020) 66 (19 February 2020). The European strategy for data aims at achieving a single market for data though a variety of means, including the enacting of sector specific legislation and the regulation of European Data Spaces. Within the framework of this strategy, two legislative proposals have been advanced, namely, the Data Act and the Data Governance Act. The latter became EU Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European Data Governance and Amending Regulation (EU) 2018/1724 [2022] OJ L152/1 (Data Governance Act).

¹³⁶ *ibid.*

¹³⁷ Kickbusch and others (n 21) 1736.

¹³⁸ On the human rights’ relevance of data solidarity, see *ibid* 1751.

¹³⁹ Editorial, ‘Can Digital Technologies Improve Health?’ (2021) 398(10312) *The Lancet* 1663.

Finally, some other specific proposals to overcome biases, errors, and the black box problem include incentivising AI developers to design devices for LMIC and increasing the transparency about training context.¹⁴⁰

5 Conclusion

This chapter provided an overview of AI technologies for health and the promises they hold for the improvement of human health at the individual and collective level. To accomplish the full potential of AI for health, some human rights challenges posed by these technologies must be tackled first, including issues of inaccuracy, unexplainability and opacity, privacy and data protection, and cybersecurity. Although a comprehensive solution to these issues may not be the best policy approach, a combination of sectoral and transdisciplinary solutions (including law and ethics) with a robust human rights framework may provide the right tools to overcome the threats to the full enjoyment of the benefits derived from the use of AI in health.

¹⁴⁰ Minssen and others (n 77) 17.

PART VIII

ARTIFICIAL INTELLIGENCE AND
THIRD GENERATION RIGHTS

Artificial Intelligence and Consumer Protection Rights

*Shu Li, Béatrice Schütte, and Lotta Majewski**

1 Introduction: Consumer Protection as a Human Right?

The rise of mass production has escalated the tension between traders and consumers, making the latter suffer from severe information asymmetry.¹ The emergence of consumer rights is to correct this market imbalance. This development shall enable consumers to make free choices and to be able to better evaluate the benefits and risks inherent to their choices.² Consumers are thus expected to influence the price and products through their corrected behaviour, and their relationship with traders could be balanced.³ Consumers are specifically protected in a variety of domains, ranging from commercial practices and contractual terms to product safety and access to justice.⁴

As abusive market behaviour is increasingly encroaching the field of consumer protection, the importance of consumer protection has been underscored within numerous international documents.⁵ Consumer protection was explicitly recognised for the first time in the United Nations Guidelines for Consumer Protection (UNGCP), adopted by the UN General Assembly (UNGA) in 1985.⁶ It clarified in several provisions that consumer protection can be strengthened by national

* This research has been conducted with the help of project funding granted by the Academy of Finland, decision number 330884 (2020). The authors would like to thank the editors as well as the entire team of the Legal Tech Lab at University of Helsinki for valuable input.

¹ Andrew F Daughety and Jennifer F Reinganum, ‘Economic Analysis of Products Liability: Theory’ in Jennifer H Arlen (ed), *Research Handbook on the Economics of Torts* (Edward Elgar 2013).

² Keith N Hylton, ‘The Law and Economics of Products Liability’ (2012) 88 *Notre Dame Law Review* 2457.

³ Bastian Schüller, ‘The Definition of Consumers in EU Consumer Law’ (2012) European Consumer Protection 123.

⁴ Geraint Howells, Christian Twigg-Flesner, and Thomas Wilhelmsson, *Rethinking EU Consumer Law* (Taylor & Francis 2017).

⁵ Iris Benöhr and Hans-Wolfgang Micklitz, ‘Consumer Protection and Human Rights’ in Geraint Howells, Iain Ramsay, and Thomas Wilhelmsson (eds), *Handbook of Research on International Consumer Law* (2nd edn, Edward Elgar 2018) 16.

⁶ United Nations Conference on Trade and Development (UNCTAD), ‘United Nations Guidelines for Consumer Protection (UNGCP)’ (2016) <https://unctad.org/system/files/official-document/ditccp_lpmisc2016d1_en.pdf>.

legislation. The UNGCP has notable influence, although it is not a binding legal document. This international endorsement later had a great impact on regions and nations in acknowledging consumer protection as a fundamental right. Several countries across the world started to include consumer protection into their constitutions.⁷ At the EU level, article 38 of the Charter of Fundamental Rights (CFR), which became binding with the adoption of the Lisbon Treaty in 2009, states that a high level of consumer protection shall be ensured. From this perspective, the necessity to protect consumers as a vulnerable group of humans is explicitly recognised.⁸

In view of the developments shown above, one may raise the question whether consumer protection falls into the scope of human rights. From a conceptual perspective, literature has argued that consumer rights hold the attributes of being a kind of human rights, which are: universally wide recognition, improving the well-being of humans, and protection against powerful governments.⁹ Therefore, the discussion on the policy development and the conceptual delineation has indicated that consumer protection per se ‘shows elements of a new generation of fundamental rights’; thus going beyond the sentiment that consumers as vulnerable actors in the digital age are already protected by a number of dissimilar human rights, such as the right to life and personal security, the right to non-discrimination, and so on.¹⁰

The recent technological development represented by artificial intelligence (AI) is a double-edged sword to consumer protection. On the one hand, it seems to promise a bright future to enhancing consumer welfare in numerous ways. On the other hand, the technology can unconsciously manipulate the behaviour of consumers and place their fundamental rights at risk. Against this background, protecting consumers in the era of AI has been the tenet of ongoing policymaking.

The European Union (EU) draft regulation on AI (AI Act) sets out to ensure the right enshrined in article 38 of the CFR alongside more ‘traditional’ fundamental rights, such as the right to human dignity.¹¹ Policy papers such as the White Paper on Artificial Intelligence mention consumer rights and consumer protection next to human rights.¹² In the explanatory memorandum to the AI Act, the European Commission (‘the Commission’) further states that the restrictions imposed by the proposal are justified to ‘ensure compliance with overriding reasons

⁷ Such as in Brazil, Switzerland, and Spain. See Benöhr and Micklitz (n 5) 21.

⁸ ibid 22.

⁹ Sinai Deutch, ‘Are Consumer Rights Human Rights’ (1994) 32 Osgoode Hall Law Journal 537.

¹⁰ Benöhr and Micklitz (n 5) 31.

¹¹ European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts’ (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD), part 3.5, Fundamental Rights, Recital 28 <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

¹² European Commission, ‘White Paper on Artificial Intelligence: A European Approach to Excellence and Trust’ COM(2020) 65 final, 10, 13.

of public interest such as health, safety, consumer protection and the protection of other fundamental rights.¹³ This implies at least a tendency to attribute consumer protection a human-right-like status.

Against this background, this chapter will not offer an answer to the question of whether or not we shall facilitate the recognition of consumer protection per se as a human right in the era of AI. Instead, our discussion sheds light on how the use of AI can be an enabler or destroyer of consumer protection from a human rights perspective. In section 2, we argue that AI can be utilised in various ways to improve consumer welfare. In section 3, we analyse how the widespread application of AI can disadvantage consumers and violate human rights. Several concrete approaches to enhance the human rights of consumers in the era of AI are outlined in section 4.

In order to better understand how consumers can be disadvantaged with the use of AI, our analysis will refer to the dilemma that the incumbent legal framework fails to tackle. Specifically, the EU will be used as an exemplary jurisdiction for interpretation. There are two main reasons for this choice: First, the EU law has a profound tradition in light of consumer protection, so we are able to have a closer look at how existing law fails to react to the threats posed by AI. Second, the EU is a forerunner in terms of AI regulation. Having issued the first concrete legislative proposals to regulate AI and related issues such as digital services and digital markets, the EU provides us with approaches towards protecting the fundamental rights of consumers in the era of AI. However, where necessary, we will also include examples from other legal systems and from international organisations.

2 AI as a Force for Good in Consumer Protection

In this section, we argue that, if managed appropriately, AI has the potential to enhance consumer protection and improve consumer welfare. Several examples are given to echo this proposition.

Many AI-driven products have been designed to assist consumers in their everyday lives.¹⁴ They can help consumers to make more informed and rational choices, and they can provide substantial assistance in cases of different special needs such as impaired vision, reduced mobility, or cognitive disabilities.¹⁵ AI can support the implementation of consumer protection law. In this context, it can help detecting law infringements or assessing compliance.¹⁶ In recent years, AI-driven

¹³ AI Act (n 11) 11.

¹⁴ Jeannie Marie Paterson and Yvette Maker, 'AI in the Home: Artificial Intelligence and Consumer Protection Law' in Ernest Lim and Phillip Morgan (eds), *The Cambridge Handbook of Private Law and Artificial Intelligence* (CUP 2022) 2.

¹⁵ *ibid* 6.

¹⁶ Giovanni Sartor, 'New Aspects and Challenges in Consumer Protection' (2020) 27 <[www.eropa.eu/RegData/etudes/STUD/2020/648790/IPOL_STU\(2020\)648790_EN.pdf](http://www.eropa.eu/RegData/etudes/STUD/2020/648790/IPOL_STU(2020)648790_EN.pdf)>.

applications have been developed to remove obstacles to consumer protection. The most common and well known ones in this regard are likely ad-blockers as well as anti-spam and anti-phishing systems.¹⁷ Price comparison websites help consumers find the best offer for the item they look to purchase.¹⁸ Nowadays, there are even augmented reality powered apps that can help consumers compare prices of products advertised in offline outlets like marketing pamphlets in newspapers.¹⁹ Another example of beneficial AI is software developed to help consumers detect potentially unfair contractual clauses.²⁰ In fact, consumers often do not read terms of services online, arguing that the documents are overwhelming. Studies have indicated that only reading privacy policies would already take about 200 hours per year. In any event, consumers have no means to influence these terms of service as they are unilaterally determined by the respective business operator.²¹ In the course of a project carried out at the European University Institute and the University of Modena, researchers developed a program called CLAUDETTE, which shall provide legal assessment of online consumer contracts.²² Additionally, researchers are developing tools that assist consumers in assessing the security and privacy measures of websites, which is particularly useful with respect to online shopping as they can make an informed choice whether or not they want to purchase goods from a certain trader.²³

Certain technologies such as blockchain can also be applied to enhance product safety and product quality and to protect against counterfeit products. Encryption used by this technology enhances protection against manipulation. At the same time, blockchain technology can help improve the traceability of products, which would contribute to more efficient product recalls. Traceability of products also allows the consumer to see the actual origins of the product or its components.²⁴ Further, sensors can monitor how products are used by consumers and how they perform. By those means, manufacturers can obtain crucial information, for instance regarding whether a product needs to be repaired or updated. This predictive maintenance could prevent product failures and accidents resulting therefrom.²⁵ Already at the pre-marketing stage, AI-enabled data collection

¹⁷ Marco Lippi and others, 'Consumer Protection Requires Artificial Intelligence' (2019) 1 *Nature Machine Intelligence* 168.

¹⁸ C Thorun and J Diels, 'Consumer Protection Technologies: An Investigation into the Potentials of New Digital Technologies for Consumer Policy' (2020) 43 *Journal of Consumer Policy* 182.

¹⁹ *ibid*.

²⁰ HW Micklitz, P Palka, and Y Panagis, 'The Empire Strikes Back: Digital Control of Unfair Terms of Online Services' (2017) 40 *Journal of Consumer Policy* 367.

²¹ Marco Lippi and others, 'CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service' (2019) 27 *Artificial Intelligence and Law* 118.

²² Thorun and Diels (n 18) 185; Lippi and others (n 17) 168.

²³ Thorun and Diels (n 18) 183.

²⁴ *ibid* 181.

²⁵ Office for Product Safety and Standards (OPSS), 'Study on the Impact of Artificial Intelligence on Product Safety, Final Report' (December 2021) 31 <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1077630/impact-of-ai-on-product-safety.pdf>.

processes used during manufacturing can help prevent mass product recalls.²⁶ Connected products further allow direct communication between producer and consumer, facilitating more efficient warnings or even the fixing of the problem in question by software updates or similar means.²⁷

New technologies can also help facilitate consumers' access to justice.²⁸ One example worth mentioning is online dispute resolution (ODR). There is no comprehensive definition of ODR yet; however, the different available mechanisms share the implementation of technology in order to increase efficiency in conflict management.²⁹ The combination of e-commerce platforms with mechanisms of ODR and payment systems could help consumers to enforce their rights without having to resort to costly and time-consuming litigation in court.³⁰ Particularly in cross-border e-commerce disputes, the traditional means of solving conflicts are often disproportionate in terms of time and cost compared to the value of the goods in question.³¹ This often results in consumers not getting any form of refund or compensation. Online marketplace, eBay, was one of the first businesses to implement ODR mechanisms. The so-called eBay Resolution Centre resolves about 60 million disagreements between e-commerce parties every year.³² Nowadays, also the Commission provides an ODR platform through which customers can solve issues with traders, aiming at increasing safety and fairness in online shopping.³³ Automated claims can also be an efficient tool, particularly in clear-cut cases. As an example, dedicated platforms in the EU allow consumers to easily file their passenger right claims³⁴ and assess the viability for a success fee when their flight is cancelled or delayed.³⁵

²⁶ *ibid* 32. See also European Commission, 'Report on the Safety and Liability Implications of AI, the Internet of Things and Robotics' COM(2020) 64 final, 3.

²⁷ *ibid*, European Commission.

²⁸ See the chapter by Helga Molbaek-Steenig and Alexandre Quemyn in this volume.

²⁹ Riikka Koulu, 'Blockchains and Online Dispute Resolution: Smart Contracts as an Alternative to Enforcement' (2016) 13 SCRIPTed 40, 42.

³⁰ *ibid* 47.

³¹ Constantina Sampani, 'Online Dispute Resolution in E-Commerce: Is Consensus in Regulation UNCITRAL's Utopian Idea or a Realistic Ambition?' (2021) 30(3) Information & Communications Technology Law 235, 236.

³² Jeremy Barnett and Philip Treleaven, 'Algorithmic Dispute Resolution: The Automation of Professional Dispute Resolution Using AI and Blockchain Technologies' (2018) 61(3) The Computer Journal 399, 404.

³³ European Commission, 'Online Dispute Resolution' <<https://ec.europa.eu/consumers/odr/main/?event=main.home.howitworks>>.

³⁴ European Parliament and Council Regulation (EC) 261/2004 of 11 February 2004 establishing common rules on compensation and assistance to passengers in the event of denied boarding and of cancellation or long delay of flights [2004] OJ L6/1 sets out that air passengers are entitled to compensation in certain cases of denied boarding, flight cancellations, or long delays.

³⁵ Thorun and Diels (n 18) 185.

3 AI as a Force for Bad in Consumer Protection

However, the influence of AI on consumers is not always as positive as we depicted in section 2. The use of AI in various business sectors can put human rights at risk as well. By processing large amounts of data, AI can significantly unbalance the relationship between traders and consumers. In this part, we will analyse how the use of AI in commercial practice has an impact on the fundamental rights of consumers. Three facets will be focused on. Section 3.1 deals with how data-driven AI technologies can be used by traders to manipulate consumers and thereby restrict their personal autonomy. Section 3.2 focuses on how digital technologies have an impact on product safety and thus on consumer rights. Section 3.3 examines immaterial harm caused by AI, using discrimination as an example.

3.1 Personal Autonomy

Personal autonomy per se is not a kind of human right under international human rights law (IHRL). However, it has been a crucial prerequisite for human dignity as well as the safeguard against the violation of other human rights. Without personal autonomy, consumers cannot make free choices.³⁶ Consumers should have a clear idea of the benefits and risks of a product or service, and make their choice based on that.³⁷ Otherwise, they could be exposed to material and immaterial harms. Therefore, autonomy serves as the starting point in the discussion on how AI poses a risk to the fundamental rights of consumers.

Experiments have found the existence of ‘algorithmic decision-making (ADM) aversion’, indicating that people are more sensitive to ADM errors despite its stability being similar to human decisions.³⁸ By applying complex analytical and predictive algorithms driven by AI, traders can detect the correlation between the data collected from consumers and their behaviour. Traders can not only recognise the consumers’ established traits and preferences but also some psychological vulnerabilities and cognitive biases.³⁹ Hence, deploying AI in business practice vests traders with a way to manipulate the consumers’ behaviour.⁴⁰ The opacity and

³⁶ Hans-Wolfgang Micklitz, Norbert Reich, and Peter Rott, *Understanding EU Consumer Law* (Intersentia 2009) 21–26.

³⁷ European Consumer Organisation (BEUC), ‘EU Consumer Protection 2.0: Protecting Fairness and Consumer Choice in a Digital Economy’ (2022) <www.beuc.eu/publications/eu-consumer-protection-20-protecting-fairness-and-consumer-choice-digital-economy/html>.

³⁸ Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey, ‘Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err’ (2015) 114 *Journal of Experimental Psychology: General* 114.

³⁹ Natali Helberger, ‘Profiling and Targeting Consumers in the Internet of Things—A New Challenge for Consumer Law’ in R Schulze and D Staudenmayer (eds), *Digital Revolution: Challenges for Contract Law in Practice* (Nomos 2016) 135.

⁴⁰ Ryan Calo, ‘Digital Market Manipulation’ (2013) 82 *George Washington Law Review* 995; and Daniel Susser, Beate Roessler, and Helen Nissenbaum, ‘Technology, Autonomy, and Manipulation’ (2019) 8(2) *Internet Policy Review* 1.

complexity of an ADM mechanism may nudge consumers to behave in a way they would not have, had they fully understood its pattern. Ultimately, the autonomy of consumers can be significantly undermined.⁴¹

The negative impact of AI on manipulating consumers can be observed in several contexts.⁴² One typical application is targeted and personalised advertising.⁴³ In the information society, a two-sided market is formed by online service providers.⁴⁴ They establish various online service infrastructures, such as online platforms, online payments, cloud services, and so on. While these services are free to consumers, the advertising avenue paid by advertisers and traders further backs the business model.⁴⁵ In other words, online intermediaries provide for a market to enhance the interdependence of traders and consumers, in which online records from consumers constitute a commodity that advertisers and traders pay for. By accessing the data and utilising AI-driven analytical tools to profile consumers, traders can distribute tailored ads to consumers.⁴⁶ Targeted advertising is a double-edged sword to consumers. On the one hand, it can reduce transaction costs for consumers since relevant contents could be displayed to them without great searching effort. On the other hand, consumers could be nudged to make undesirable choices. Digital services have been designed to facilitate data collection, so that traders can capture every footprint (eg a click on a web page) of consumers, exploiting their prejudice and trigger negative feelings.⁴⁷ From this latter perspective, the information gap between traders and consumers is considered to be even larger, since traders not only hold better information of products or services, but also they even know consumers better than consumers themselves do.⁴⁸ Data protection regulation has been a key approach to addressing exploitation through asymmetry of information, the premise being that requiring informed consent from consumers to use their data lessens the traders' opportunity for exploitation.⁴⁹ While data protection concerns the right to privacy, it often complements the goals of consumer law, although unanswered questions remain regarding their

⁴¹ Cass R Sunstein, 'Sludge and Ordeals' (2018) 68 Duke Law Journal 1843.

⁴² See eg Agnieszka Jabłonowska and others, 'Consumer Law and Artificial Intelligence: Challenges to the EU Consumer Law and Policy Stemming from the Business' Use of Artificial Intelligence-Final Report of the ARTSY Project' (EUI Department of Law Research Paper 2018/11, 2018) <<http://hdl.handle.net/1814/57484>>.

⁴³ Sophie C Boerman and others, 'Online Behavioral Advertising: A Literature Review and Research Agenda' (2017) 46 Journal of Advertising 363.

⁴⁴ Jean-Charles Rochet and Jean Tirole, 'Two-sided Markets: A Progress Report' (2006) 37(3) RAND Journal of Economics 645.

⁴⁵ Sartor (n 16) 12.

⁴⁶ ibid.

⁴⁷ Natali Helberger and others, 'EU Consumer Protection 2.0: Structural Asymmetries in Consumer Markets' (2021) <www.beuc.eu/publications/beuc-x-2021-018_eu_consumer_protection.0_0.pdf>.

⁴⁸ Sartor (n 16) 14.

⁴⁹ See eg Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (GDPR), Preamble 32, 39.

interplay.⁵⁰ Data protection and the right to privacy encompass broad societal interests that are larger in scope than the interests of consumer rights.⁵¹ However, some argue that consumer law offers more effective remedies to privacy violations than data protection or competition law.⁵²

3.2 Product Safety and Product Liability

Everyone has the right to ensure their body and property are free from being harmed. The rights to life, personal security, and protection of property are widely recognised by international human rights law (IHRL).⁵³ In modern society, people cannot live without consumption. Products, however, are not always of perfect quality when they are placed on the market. Unsafe products, accordingly, may adversely affect the enjoyment of life and personal security. Unsafe products lead to high costs for consumers and the society since they cause injuries or premature deaths amounting to about 76 billion euros per year. It is estimated that consumers suffer a detriment of roughly 19 billion euro due to purchased unsafe consumer products, which they would not have purchased had they known that they were unsafe.⁵⁴ Consumer safety is further challenged by the increasing use of e-commerce. For instance, since 2002, the percentage of Europeans shopping online increased from 9 per cent back then to over 70 per cent nowadays, and 20 per cent of the companies in the EU are selling online. This development was additionally pushed by the COVID-19 pandemic and the lockdowns imposed in most EU member states, which led to an additional increase in online sales alongside a drop in retail sales.⁵⁵

While AI and similar technologies can improve the safety of products,⁵⁶ their specific inherent characteristics like autonomy, data dependency, connectivity, or opacity can have a negative impact on the safety of consumer products.⁵⁷ For instance, connectivity can lead to increased vulnerability to cyberattacks while

⁵⁰ Frederick Borgesius, Natali Helberger, and Reyna Agustin, ‘The Perfect Match? A Closer Look at the Relationship Between EU Consumer Law and Data Protection Law’ (2017) 54 Common Market Law Review 1427.

⁵¹ *ibid* 1463.

⁵² Maureen Ohlhausen and Alexander Okuliar, ‘Competition, Consumer Protection, and the Right (Approach) to Privacy’ (2015) 80 Antitrust Law Journal 121, 155.

⁵³ For the right to life and personal security, see eg art 12 of the Universal Declaration of Human Rights (UDHR).

⁵⁴ European Commission, ‘Commission Staff Working Document: Impact Assessment Accompanying the Document “Proposal for a Regulation of the European Parliament and of the Council on General Product Safety, Amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council, and Repealing Council Directive 87/357/EEC and Directive 2001/95/EC of the European Parliament and of the Council”’ SWD(2021) 168 final, 10–11 (European Commission, ‘Impact Assessment SWD(2021)').

⁵⁵ *ibid* 14.

⁵⁶ See section 2.

⁵⁷ European Commission (n 26) 4; OPSS (n 25) 35.

autonomy may result in unpredicted outcomes, and opacity generates obstacles to understand the decision-making process of an AI system.⁵⁸ AI-based products can be a threat to consumers by either bringing new risks to their health and safety or changing the way the existing risks could materialise.⁵⁹

The risk increased by AI further derives from the ambiguity of legal rules, which fail to echo the damages and leave consumers a difficult judiciary access to be remedied. Gaps in the regulatory regime may arise when consumer products with AI features learn and make decisions autonomously after being placed on the market.⁶⁰ Existing rules on product safety and product liability have been enacted before products connected to the internet of things and other sophisticated AI-driven devices existed or were anticipated at all.⁶¹ With increasing digitalisation, the line between products and services becomes increasingly blurry. Further, it is not evident whether standalone software based on AI qualifies as a product.⁶² The applicability of existing product safety regulations to new technologies, and particularly AI-related devices, is not entirely clear. Particularly more sophisticated AI applications challenge the concepts and definitions enshrined in conventional product safety and product liability legislation.⁶³ For instance, with regard to European law, the General Product Safety Directive (GPSD) in its current version does not mention directly that digital features like AI can have an impact on product safety.⁶⁴

The number of connected devices is constantly increasing. In 2019, there were an estimated 14.2 billion connected devices worldwide. By 2025, it is expected that in Europe alone there will be almost 5 billion connected devices.⁶⁵ To echo the technological development, action is being taken both at the EU level and in third countries. The Commission published in June 2021 a proposal for a Regulation on general product safety as well as a Staff-Working Document (SWD) assessing the impact of a revision of the GPSD. The aim of this recent proposal is to strengthen the protection of safety and health for European consumers and to guarantee their right to information.⁶⁶ Regarding e-commerce, the Commission finds in its SWD that the GPSD in its current version does not provide for sufficient online market

⁵⁸ European Commission (n 26) 5–9.

⁵⁹ European Commission, ‘Impact Assessment SWD(2021)’ (n 54) 12, 13.

⁶⁰ OPSI (n 25) 55.

⁶¹ Gabriele Mazzini, ‘A System of Governance for Artificial Intelligence through the Lens of Emerging Intersections between AI and EU Law’ (2019) 7. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3369266>.

⁶² *ibid.*

⁶³ *ibid* 8; European Commission, ‘Impact Assessment SWD(2021)’ (n 54) 12, 13.

⁶⁴ European Commission, ‘Impact Assessment SWD(2021)’ (n 54) 12.

⁶⁵ *ibid* 11.

⁶⁶ European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council on General Product Safety, Amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council, and Repealing Council Directive 87/357/EEC and Directive 2001/95/EC of the European Parliament and of the Council’ COM(2021) 346 final, 14.

surveillance. This affects particularly the so-called non-harmonised products, which are not covered by specific EU legislation. Surveillance becomes even more difficult on online marketplaces where economic operators can reach an indefinite number of consumers. In addition, an increasing number of products are purchased directly from economic operators established outside the EU.⁶⁷ This makes it even more difficult to monitor the supply chains.

The increasing use of AI also challenges the applicability of existing rules on product liability as many AI systems will still qualify as products under the scope of the framework.⁶⁸ For example, it could become more difficult for consumers to obtain compensation based on existing product liability regimes. So far, for instance, under the EU Product Liability Directive (PLD),⁶⁹ the producer is liable if the product was defective at the time it was placed on the market.⁷⁰ However, products with AI features are subject to changes during their life cycle. A defect inflicted by a faulty update is thus excluded by the PLD. Digitalisation also poses obstacles to consumers regarding the burden of proof. Pursuant to article 4 of the PLD, injured parties must prove the defect, the damage, and the causal link between both. As previously mentioned, without reasonable information disclosure measures, consumers do not have the same knowledge that businesses have, and to acquire the necessary insights would require the consumer to invest time and money. This might deter the consumer from filing a damages claim under the product liability framework.⁷¹ To address the above-mentioned issues in relation to product liability, the European Commission issued in September 2022 a proposal for a revised Product Liability Directive.⁷² Pursuant to article 8, in the future a potentially liable economic operator shall be obliged to disclose relevant evidence, if the court orders them to do so upon the request of the claimant—provided that the latter had presented facts and evidence supporting the plausibility of their claim. The draft further contains alleviations to the burden of proof in favour of the injured party in article 9 in the form of rebuttable presumptions of defectiveness, or of a causal link between the defect and the damage. The suggested provisions shall help to mitigate the information asymmetry between the injured party and the economic operator.

⁶⁷ European Commission, 'Impact Assessment SWD(2021)' (n 54) 15.

⁶⁸ Which is particularly the case for so-called smart products. See eg Sebastian Lohsse, Reiner Schulze, and Dirk Staudenmayer, 'Liability for Artificial Intelligence' in Sebastian Lohsse, Rainer Schulze, and Dirk Staudenmayer (eds), *Liability for Artificial Intelligence and the Internet of Things* (Hart 2019) 15–16; see also Béatrice Schütte, Lotta Majewski, and Katri Havu, 'Damages Liability for Harm Caused by Artificial Intelligence – EU Law in Flux' (2021) Helsinki Legal Studies Research Paper No 69 1, 8.

⁶⁹ Council Directive 85/374/EEC of 25 July 1985 on the Approximation of the Laws, Regulations and Administrative Provisions of the Member States Concerning Liability for Defective Products (Product Liability Directive/PLD) [1985] OJ L210/29.

⁷⁰ See PLD, art 6; see also eg Daily Wuyts, 'The Product Liability Directive: More than two Decades of Defective Products in Europe' (2014) 5 Journal of European Tort Law 21.

⁷¹ See eg European Commission (n 26) 12; see also Schütte, Majewski, and Havu (n 68) 23, 26.

⁷² Proposal for a Directive of the European Parliament and of the Council on liability for defective products, COM(2022) 495 final.

In addition, problems arise as many AI systems may fall outside the scope of the current legal regimes. As regards the definition of a product in the digital age, from an EU perspective, the proposal for a revised PLD clarifies in its article 4(1) that software, also when supplied as standalone software, can be a product. Previously, the Commission had already suggested a new definition of a ‘product’ in the Proposal for a Regulation on General Product Safety: pursuant to article 3(1), a product is ‘any item, interconnected or not to other items, supplied or made available, whether for consideration or not, in the course of a commercial activity including in the context of providing a service—which is intended for consumers or can, under reasonably foreseeable conditions, be used by consumers even if not intended for them’. While this definition makes a clear reference to interconnected devices and uses the term ‘item’ instead of ‘movable’ as used in article 2 of the current PLD even this definition failed to clarify whether ‘item’ refers to tangibles only or whether it includes intangibles such as standalone software or data. However, the latest proposal on product liability provides clarification in this context. In any event, the definitions of the term ‘product’ will need to be aligned between the future PLD and the GPSR. For a long time, it has been asked whether the distinction between tangibles and intangibles was still relevant.⁷³ The supporting argument is that consumers in the EU use intangible goods every day and should thus be entitled to compensation. It is hard to distinguish a situation where a product as such causes damage from the situation where an integrated, yet separable, software program in the product causes damage.⁷⁴

Also from a United States (US) perspective, it is acknowledged that interactions with consumer-users and the related data gathering can affect the decision-making patterns of AI systems and consequentially lead to unforeseen injuries.⁷⁵ Actual case law addressing liability for harmful products with AI features is scarce. The best-known case is probably the crash of a self-driving Uber vehicle in Arizona, in which a pedestrian was killed.⁷⁶ The related civil proceedings ended with a settlement concluded between the company and the bereaved.⁷⁷ Regarding the question whether online games are products, the United States District Court, Western District of Washington defined in March 2022 a ‘product’ as ‘any object possessing intrinsic value, capable of delivery either as an assembled whole or as a component part or parts, and produced for introduction into trade or

⁷³ See eg Charlotte de Meeus, ‘The Product Liability Directive at the Age of the Digital Industrial Revolution: Fit for Innovation?’ (2019) 8 Journal of European Consumer and Market Law 151.

⁷⁴ *ibid.*

⁷⁵ See eg Greg Swanson, ‘Non-Autonomous Artificial Intelligence Programs and Products Liability: How New AI Products Challenge Existing Liability Models and Pose New Financial Burdens’ (2019) 42 Seattle University Law Review 1203.

⁷⁶ See eg BBC News, ‘Uber’s Self-Driving Operator Charged over Fatal Crash’ (BBC News, 16 September 2020) <www.bbc.com/news/technology-54175359>.

⁷⁷ See eg Tom McKay, ‘Uber Not Criminally Liable in Fatal 2018 Self-Driving Car Accident, Arizona Prosecutor Finds’ (Gizmodo, 5 March 2019) <<https://gizmodo.com/uber-not-criminally-liable-in-fatal-2018-self-driving-c-1833082126>>.

commerce'. Consequentially, online games are a software-based service.⁷⁸ Thus, as per the current approach, software is only a product when incorporated into a tangible item. Under general US Product Liability Law,⁷⁹ a manufacturer is liable for manufacturing defects, design defects, and warning defects. Particularly design defects are considered problematic under AI involvement, as both the AI and the consumer are controlling the product.⁸⁰ In this context, an important question relates to which interactions between the consumer and the AI must be foreseeable, and to what extent a manufacturer is obliged to mitigate certain interactions that might lead to pattern changes by an alternative design.⁸¹ However, there is no uniform approach to the question of foreseeability across US states, which means that harm-sufferers might face high obstacles to prove that the damage sustained was actually foreseeable, depending on the state they are based in.⁸²

In general, the discussion above shows that across jurisdictions similar, if not the same aspects in terms of product safety and product liability are perceived as challenging in regard to increasingly digitalised and interconnected consumer products.

3.3 Discrimination

Besides physical harms to health and property, AI in commercial practice can also generate immaterial harm, which brings pain and suffering to consumers or leads to a loss of chance. One noteworthy example of such immaterial harm derives from discrimination.⁸³ AI has been repeatedly reported to generate serious discriminatory outcomes when it is used for predictive policing, recidivism as well as employment.⁸⁴ This section reveals that algorithmic discrimination also exists in the business setting. The disadvantaged consumers may suffer more social inequality in light of accessing goods and services.

Consumers should not face unreasonable obstacles in their access to goods and services, based on characteristics such as gender, race, religion, wealth, and so on.⁸⁵ However, with the deployment of AI in business practice, these features might be abused. Direct discrimination may already occur even if consumers are unaware of

⁷⁸ *Penny Quiteros, Plaintiff, v INNOGAMES et al, Defendants* No C19-1402RSM (WD Wash, 30 March 2022) <<https://casetext.com/case/quiteros-v-innogames-3>>.

⁷⁹ See Restatement of US Tort Law (Third), Section 402A.

⁸⁰ Swanson (n 75) 1213.

⁸¹ *ibid* 1216.

⁸² *ibid* 1218.

⁸³ For the codification of non-discrimination rights, see eg Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR), art 14; and Charter of Fundamental Rights of the European Union [2012] OJ C326/391, art 21.

⁸⁴ See the chapter by Valentina Gulinova in this volume.

⁸⁵ ECHR, art 14.

it.⁸⁶ Bias towards consumers can emerge at any point in the course of an AI operation. First, it can occur in the process of data extraction and acquisition of training data. When the data samples used to train algorithms fail to properly represent the population, the AI system will duplicate such a bias.⁸⁷ Secondly, the features used during model building and testing are chosen by humans, so that the result to some extent can replicate and reinforce the designers' bias.⁸⁸

The typical characteristics covered by incumbent non-discrimination law are limited (eg gender, race, etc). In the era of AI, even if the identifiable protected features have been well controlled, discrimination can still occur. This is because many other new but not immediately evident proxies may be included into the patterns of algorithms to manipulate consumers.⁸⁹ The deployment of AI in business practice can lead to *indirect discrimination* on a large scale. That means, although traders apply neutral and alike criteria to all consumers, some proxies (such as postal code) are still correlated with protected characteristics. As a result, the discrimination effect can still be duplicated, even if protected characteristics have already been removed from the model.⁹⁰ From a technical perspective, such an indirect discrimination is always a sort of 'unintentional side effect' of the algorithm's use, which is beyond the consciousness of programmers.⁹¹ From a legal perspective, both direct and indirect discrimination are not allowed. The problem raised by AI is that the current non-discrimination laws may not well echo the indirect discrimination induced by algorithms, since it is often justified by the predictive accuracy of the model.⁹² What is more, neutral proxies that are measured in the model are not covered by the laws.⁹³ Even worse, it may be beyond the capacity of consumers and regulators in light of detecting or understanding how a specific proxy can be correlated with protected characteristics and result in discrimination.⁹⁴ The analysis thereby reveals that AI has enhanced discrimination, which cannot be easily overcome under the incumbent anti-discrimination laws.

⁸⁶ Amy J Schmitz, 'Secret Consumer Scores and Segmentations: Separating Haves from Have-Nots' (2014) Michigan State Law Review 1411.

⁸⁷ Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671.

⁸⁸ Janneke Gerards and Raphaële Xenidis, 'Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Non-discrimination Law' (2020) <<https://op.europa.eu/en/publication-detail/-/publication/082f1dbc-821d-11eb-9ac9-01aa75ed71a1>>.

⁸⁹ Frederik Zuiderveen Borgesius, 'Discrimination, Artificial Intelligence, and Algorithmic Decision-Making' (CoE, 2018) <<https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73>>.

⁹⁰ Philipp Hacker, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law' (2018) 55 Common Market Law Review 1143.

⁹¹ Barocas and Selbst (n 87).

⁹² Hacker (n 90).

⁹³ Janneke Gerards and Frederik Zuiderveen Borgesius, 'Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence' (2022) 20 Colorado Technology Law Journal 1–55.

⁹⁴ Frederik Zuiderveen Borgesius, 'Strengthening Legal Protection Against Discrimination by Algorithms and Artificial Intelligence' (2020) 24 International Journal of Human Rights 1572, 1584.

It must be noted that delegating the decision on market access to algorithms may harm consumers in some sectors more than in others. For instance, financial institutions may use AI to score consumers to ‘better’ assess their credit eligibility.⁹⁵ This has likely greater impact on consumers than discrimination imposed by chatbots utilised in customer service. Therefore, the existence of the sectoral disparity in light of AI discrimination manifests that the type and level of regulation might be different across diverse sectors.⁹⁶

The use of AI also makes price discrimination a contentious issue. In general, price discrimination means that the same product is priced differently for different consumers.⁹⁷ From an economics perspective, the price that people are willing to pay for a certain product can differ among consumers. The goal of customised pricing is to optimise the producer’s or trader’s gain.

Price discrimination, which is based on consumer’s willingness to pay, is in nature regarded as a business strategy and it is not explicitly forbidden by consumer law.⁹⁸ To segment the market and exploit consumers, producers need sufficient information to understand consumer behaviour without significant expenses. This is also exactly the place where AI can wedge.⁹⁹ By tracing personal footprints with the application of advanced algorithms, there can be no doubt that traders can better understand and exploit the preference of consumers with the help of big data and then charge them a personalised price.¹⁰⁰ By having more information about the dynamics of supply and demand, the traders that are empowered by algorithms can react quickly in the market and thus exploit more advantages than those who are not.¹⁰¹ However, personalised pricing can turn to a ‘race to the bottom’, if specific attributes of consumers are (un)consciously included as proxies for ADM.¹⁰² By doing so, price discrimination can be closely linked to indirect discrimination, and AI would serve an accelerator for this connection. In this regard, indirect discrimination may be veiled in a manner of price discrimination in the era of AI, especially when the proxies used for measurement are correlated

⁹⁵ Tal Z Zarsky, ‘Understanding Discrimination in the Scored Society’ (2014) 89 Washington Law Review 1375. According to the new EU AI Regulation, social scoring by government is forbidden. However, scoring consumers in business practice is permitted.

⁹⁶ Max N Helveston, ‘Consumer Protection in the Age of Big Data’ (2015) 93 Washington University Law Review 859.

⁹⁷ Joost Poort and Frederik J Zuiderveen Borgesius, ‘Does Everyone Have a Price? Understanding People’s Attitude Towards Online and Offline Price Discrimination’ (2019) 8(1) Internet Policy Review 1.

⁹⁸ See eg European Commission, ‘Guidance on the Implementation/Application of Directive 2005/29/EC on Unfair Commercial Practices’ SWD(2016) 163 final, 134; see also Sartor (n 16) 36.

⁹⁹ Axel Gautier, Ashwin Ittoo, and Pieter Van Cleynenbreugel, ‘AI Algorithms, Price Discrimination and Collusion: A Technological, Economic and Legal Perspective’ (2020) 50 European Journal of Law and Economics 405.

¹⁰⁰ Oren Bar-Gill, ‘Algorithmic Price Discrimination: When Demand Is a Function of Both Preferences and (Mis) Perceptions’ (2018) Harvard Public Law Working Paper No 18-32, 1.

¹⁰¹ Mateusz Grochowski and others, ‘Algorithmic Price Discrimination and Consumer Protection: A Digital Arms Race?’ (2022) Technology and Regulation 36.

¹⁰² ibid.

with protected characteristics. Considering its potentially large-scale manipulative effect, the benefit reaped by traders at the expense of consumer welfare is under scrutiny. Recent developments in consumer protection have started to respond to price discrimination, meaning that traders may be bound by specific regulation when using the strategy in their business.¹⁰³ In addition, price discrimination may be tackled by data protection law, requiring traders to have consumers' consent on processing their data for price personalising purpose.¹⁰⁴ As AI is deployed in large scale in business practice, it is advisable to specify the permissibility of price discrimination according to the sector.¹⁰⁵ Otherwise, it will be to the detriment of consumers. For example, in the insurance industry, researchers have identified that while AI can help insurers to better differentiate the risk inherent to various insured parties, it may as well generate a side effect of reinforcing social inequality by making insurance unaffordable to the poor.¹⁰⁶

The previous discussion has shown that differentiation *per se* is not a bad thing. However, efficiency cannot overwhelm fairness in all aspects, especially when the application of AI in business practice can reinforce inequality and restrict market access.

4 Protecting Consumers in the Era of AI: Approaches and Challenges to Improve the Current Status Quo

The discussion so far has indicated that the fundamental rights of consumers can be violated in ways that they do not even realise. This section will discuss the legal approaches to protect the fundamental rights of consumers in the era of AI.

Behavioural studies have shown that if the logic and consequences of ADM can be explained well to people, there is less ADM aversion.¹⁰⁷ Therefore, transparency is of great importance to consumers to enhance self-autonomy and to correct the increasing information asymmetry in business practice.¹⁰⁸ The current accessibility of large quantities of information does not mean that consumers are aware

¹⁰³ Eg according to article 6(1) of the European Parliament and Council Directive 2011/83/EU of 25 October 2011 on Consumer Rights [2011] OJ L304/64 (hereafter Consumer Rights Directive 2011/83/EU) traders are obliged to inform consumers 'where applicable, that the price was personalised on the basis of automated decision making'.

¹⁰⁴ Frederik Zuiderveen Borgesius and Joost Poort, 'Online Price Discrimination and EU Data Privacy Law' (2017) 40 Journal of Consumer Policy 347.

¹⁰⁵ Sartor (n 16).

¹⁰⁶ Zuiderveen Borgesius (n 94) 1584.

¹⁰⁷ Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey, 'Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If they Can (even slightly) Modify Them' (2018) 64 Management Science 1155.

¹⁰⁸ Andrew D Selbst and Solon Barocas, 'The Intuitive Appeal of Explainable Machines' (2018) 87 Fordham Law Review 1085.

of the consequences.¹⁰⁹ For instance, consent is considered essential to profile consumers via AI-driven personalised advertising.¹¹⁰ However, the design can deceive consumers so that they are unaware that they are interacting with AI.¹¹¹ Improved transparency would see consumers informed in a clear way about when they are interacting with AI, when a decision is made by AI, how automated decisions are made, and how they can affect them.¹¹² Researchers and legislators have proposed a variety of measures to improve transparency in business practice.

One of the solutions is granting consumers the right to access information on the actual process of ADM. Data protection law has in this respect been endorsed as an important approach in terms of protecting fundamental rights.¹¹³ Consumers who suffer algorithmic discrimination may rely on data protection law to seek access to and explanation of the data and algorithms used for decision-making. One example is the request for a specific right to explanation of ADM, meaning that traders must explain to consumers how a decision is made by algorithms and what the consequences are.¹¹⁴ Granting consumers the right to explanation, either ex ante or ex post, is expected to balance out the information asymmetry between traders and consumers. This requirement has been implicitly reflected in recent legislation.¹¹⁵ However, many scholars argue that such a right remains largely unclear for now.¹¹⁶ The implementation of the right to explanation may also be difficult in practice. For instance, algorithms might be qualified as other legal entitlements such as trade secrets or intellectual property.¹¹⁷ The exposure to algorithms may place privacy at

¹⁰⁹ David Bawden and Lyn Robinson, ‘The Dark Side of Information: Overload, Anxiety and Other Paradoxes and Pathologies’ (2009) 35 *Journal of Information Science* 180.

¹¹⁰ GDPR, art 21(2).

¹¹¹ Midas Nouwens and others, ‘Dark Patterns After the GDPR: Scraping Consent Pop-Ups and Demonstrating their Influence’ (*Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020) <<https://dl.acm.org/doi/10.1145/3313831.3376321>>.

¹¹² Jessica Fjeld and others, ‘Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI’ (*Berkman Klein Center Research Publication 2020-1*, 2020) <<https://dx.doi.org/10.2139/ssrn.3518482>>.

¹¹³ Zuiderveen Borgesius (n 94) 1579.

¹¹⁴ Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, ‘Transparent, Explainable, and Accountable AI for Robotics’ (2017) 2 *Science Robotics* 1.

¹¹⁵ Gianclaudio Malgieri and Giovanni Comandé, ‘Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation’ (2017) 7 *International Data Privacy Law* 243.

¹¹⁶ Eg in the GDPR, the right of explanation of ADM is only explicitly mentioned in Recital 71. However, this arrangement implies that the right of explanation of ADM is not legally binding in practice. Also, it is not clear from Recital 71 what proxies are in the scope of the right. See Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’ (2017) 7 *International Data Privacy Law* 76; Lilian Edwards and Michael Veale, ‘Slave to the Algorithm: Why a Right to an Explanation is Probably Not the Remedy You Are Looking For’ (2017) 16 *Duke Law and Technology Review* 18.

¹¹⁷ Balazs Bodo and others, ‘Tackling the Algorithmic Control Crisis: The Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents’ (2017) 19 *Yale Journal of Law and Technology* 133; Gianclaudio Malgieri, ‘Trade Secrets v Personal Data: A Possible Solution for Balancing Rights’ (2016) 6 *International Data Privacy Law* 102. See also the chapter by Letizia Tomada and Raphaële Xenidis in this volume.

risk as well.¹¹⁸ Therefore, a right of explanation toward ADM may not only have a chilling effect on innovation but also collide with other fundamental rights.¹¹⁹ What is more, ADM mechanisms are likely ‘black boxes’ that are opaque to humans.¹²⁰ Even if the underlying codes and the logic of algorithms are disclosed, the explanation of particularly complex models (such as deep learning), based on non-linear models or including hidden layers, may be beyond the interpretive capacity of consumers and regulators.¹²¹ Explainability by itself can be counterproductive to the goal of predictive accuracy. Some algorithmic structures built upon neural networks are more effective in light of natural language processing and can thus elicit more predictive accuracy than others (eg linear or rule-based models), but at the same time, they have a relatively low degree of explainability.¹²² Finally, highly explainable and transparent algorithms are susceptible to cyber-attacks, which can pose additional risks to consumers. To sum up, establishing a right to full explanation is not only impossible but also undesirable. Instead, one should focus on ensuring a *sufficient* level of explanation to protect consumers in the era of AI.

Besides granting consumers the right to explanation, traders should be subject to some information duties and accountabilities regarding the algorithms deployed in their business. For example, traders have already been requested to disclose the main parameters used for products ranking¹²³ and to inform consumers whether an ADM-based price personalisation exists.¹²⁴ Considering the dominant role of online intermediary service providers regarding how they can utilise ADM to shape information flows by determining the set-up of mechanisms such as recommender system and online advertising, recent regulations also start to focus on the obligations of these parties in light of improving transparency. Taking the EU as an example, the Digital Services Act (DSA) Package adopted by the Commission, consisting of the DSA and the Digital Markets Act (DMA), addresses societal and economic concerns of the power imbalance between large online platforms and their users. The DSA obliges all providers of online intermediary services to take

¹¹⁸ Brent Daniel Mittelstadt and others, ‘The Ethics of Algorithms: Mapping the Debate’ (2016) 3 Big Data & Society 1, 6.

¹¹⁹ Danielle Keats Citron and Frank Pasquale, ‘The Scored Society: Due Process for Automated Predictions’ (2014) 89 Washington Law Review 1.

¹²⁰ Jenna Burrell, ‘How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms’ (2016) 3 Big Data & Society 1.

¹²¹ Citron and Pasquale (n 119) 1.

¹²² Burrell (n 120); Philipp Hacker and others, ‘Explainable AI Under Contract and Tort Law: Legal Incentives and Technical Challenges’ (2020) 28 Artificial Intelligence and Law 415.

¹²³ Eg article 7(4) of the European Parliament and Council Directive 2005/29/EC of 11 May 2005 Concerning Unfair Business-to-Consumer Commercial Practices in the Internal Market [2005] OJ L149/22, providing that the main parameters used for ranking products are categorised as the ‘material information’, which traders are obliged to provide. Otherwise, the business practice conducted by traders would be defined as ‘misleading omissions’. A similar requirement on disclosing main parameters to consumers can also be found in article 6 of the new Consumer Rights Directive 2011/83/EU (n 103).

¹²⁴ Consumer Rights Directive 2011/83/EU (n 103) Recital 45.

the responsibility of transparency reporting.¹²⁵ In particular, those online marketplaces that are defined as the ‘very large platforms’ must undertake incremental obligations to disclose how advertising is prioritised and targeted. They shall clearly present to the recipients of the service (such as consumers) the main parameters of recommender systems in a comprehensible manner.¹²⁶

Other typical toolkits that are expected to increase transparency are regular algorithmic impact assessment and algorithmic auditing. Such requirements have been reflected by the DSA. The ‘very large platforms’ are obliged to regularly conduct risk assessment¹²⁷ and take independent auditing,¹²⁸ the outcome of which will help these parties to adapt their design of the ADM system to mitigate the potential risk. This also enshrines the necessity of technical measures. In recent years, influenced by article 25 of the GDPR (privacy by design), transparency by design for algorithms has been increasingly discussed by scholarship.¹²⁹

The discussion so far has indicated that the measures to improve transparency have the potential to improve the protection of consumers in the era of AI, but it meanwhile does not ensure a destined elimination of all risks. Besides respecting transparency, legal rules shall also need to be adapted in a way to protect the pluralism of the choice of consumers, thus increasing the consumer’s bargaining power and lowering prices to increase consumer welfare. In this regard, the protection of consumer autonomy is also coherently related with the issue of competition.

Large entities who play a dominant role in deciding what kind of service can be accessed by consumers shall be further regulated. To this end, for example, in the EU, the DMA shall target the activities of large online platforms (‘gatekeepers’), to vest consumers with more options of services. Gatekeepers are pursuant to article 3 of the DMA providers of core platforms services having significant impact on the internal market, serving as a gateway between business users and end users and enjoying a durable position already or in the foreseeable near future. These platforms benefit from strong network effects, intermediate the majority of transactions between business users and end users and often comprehensively trace and profile their users.¹³⁰ To mitigate negative effects of the significant power imbalance, the latest version of the proposal sets out for instance that gatekeepers must not utilise personal data from users accessing the service provided by a third party

¹²⁵ European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act/DSA) and amending Directive 2000/31/EC’ (DSA) COM(2020) 825 final, 2020/0361(COD), art 23.

¹²⁶ *ibid* art 29(1).

¹²⁷ *ibid* art 26.

¹²⁸ *ibid* art 28.

¹²⁹ Riikka Koulu, ‘Crafting Digital Transparency: Implementing Legal Values into Algorithmic Design’ (2021) 8 *Critical Analysis of Law* 81.

¹³⁰ European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council on contestable and fair Markets in the Digital Sector (Digital Markets Act)’ COM(2020) 842 final, Explanatory Memorandum, 1.

when that third-party service employs the gatekeeper's platform.¹³¹ In the future, gatekeepers will not be allowed any longer to bundle services.¹³² This will also provide consumers with more options to choose from.

Human oversight is also considered a way of securing fundamental rights of consumers.¹³³ On the one hand, by granting consumers the right of overriding and contesting the decisions made by AI, the value of human autonomy is respected.¹³⁴ On the other hand, by requiring relevant parties (eg programmers, service providers, and traders) to intervene in different manners, consumers may expect a proper overall performance of an AI system.¹³⁵ Human oversight, however, might not be the panacea we imagined. The provision of overriding options may pose more physical risks to a consumer when using an AI-driven device (eg autonomous vehicles (AV)).¹³⁶ Also, as scholarship has argued, human oversight might be inhibited unless the opacity of ADM is unpacked.¹³⁷

The recent international efforts have been keen on reflecting the necessity to protect fundamental rights of consumers. Business sectors are required to take concrete measures to protect the fundamental rights of consumers. For instance, in order to adapt to the context of digital economy, the UNGCP added new principles for good business practices, with a highlight of providing consumers using electronic commerce with the same level of protection as those engaging in other forms of commerce.¹³⁸ Further, the United Nations Guiding Principles on Business and Human Rights (the UNGPs) established a general framework requiring business to observe human rights due diligence (HRDD) in light of assessing the impact on human rights.¹³⁹ As the UN Human Rights Council has reported, HRDD shall apply to all business processes, ranging from development to deployment as well as operation.¹⁴⁰ The aim of HRDD is to identify, assess, and mitigate the adverse impact of AI applications on human rights. Likewise, the Toronto Declaration also

¹³¹ Luca Bertuzzi, 'DMA: Significant Additions Made it Into the Final Text' (*EURACTIV*, 14 April 2022) <www.euractiv.com/section/digital/news/dma-significant-additions-made-it-into-the-final-text/>.

¹³² *ibid.*

¹³³ High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (Publications Office of the European Union, 2019) <<https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1>>.

¹³⁴ Mikolaj Firlej and Araz Taeihagh, 'Regulating Human Control Over Autonomous Systems' (2021) 15 Regulation and Governance 1071, 1079.

¹³⁵ Christian J Gerdes and Sarah M Thornton, 'Implementable Ethics for Autonomous Vehicles' in M Maurer and others (eds), *Autonomes Fahren* (Springer 2015) 87.

¹³⁶ Alexander Hevelke and Julian Nida-Rümelin, 'Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis' (2015) 21 Science and Engineering Ethics 619, 624.

¹³⁷ Burrell (n 120).

¹³⁸ UNCTAD (n 6 Principle (J)).

¹³⁹ United Nations, 'Guiding Principles on Business and Human Rights' (2011) <www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf>.

¹⁴⁰ UN Human Rights Council, 'The Right to Privacy in the Digital Age: Report of the United Nations High Commissioner for Human Rights' (2021) 12 <https://media.business-humanrights.org/media/documents/A_HRC_48_31_AdvancedEditedVersion.pdf>.

targets AI-based discrimination and intends to ‘draw attention to the relevant and well-established framework of international human rights law and standards’.¹⁴¹ The Declaration stresses that the state must conduct regular human rights impact assessments (HRIA) to mitigate harm induced by discrimination when AI is utilised in public sector systems.¹⁴² Private actors, in comparison, are obliged to observe comprehensive HRDD.¹⁴³

5 Conclusion

While AI has the potential to enhance consumer protection, it also poses significant risks to consumers. The analysis in this chapter shows that the increasing application of AI in commercial practice has an even more far-reaching impact on consumers, considering the manifold ways in which it can infringe consumers’ fundamental rights. More specifically, AI can manipulate the autonomy of consumers, raise safety concerns, and amplify indirect discrimination. To deal with these challenges, two tendencies have been identified in current legislative activities in light of reinforcing consumer protection. On the one hand, consumer protection has been increasingly endowed with the status as human right. This trend has been observed in various international documents. On the other hand, concrete measures have been proposed to strengthen consumers and they shall be further materialised for the purpose of protecting consumers’ fundamental rights. In this regard, principles such as transparency, fairness, and human oversight must be ensured, not only by legal rules but also by technical measures. In addition, according to the recent international recognition of consumer protection through a lens of human rights, a comprehensive human rights due diligence notion must be ensured by various business actors, covering all phases ranging from development to deployment and operation. While it is important to monitor the use of AI in business practice and to mitigate harmful effects, one should not forget that digital features have also the potential to empower consumers. Instead of painting a too dystopic picture of new technologies, it is advisable to incentivise the uptake of mechanisms that are beneficial to consumers and to make them easily accessible for everyone.

¹⁴¹ Toronto Declaration (2018), s 2 <www.torontodeclaration.org/>.

¹⁴² ibid s 30.

¹⁴³ ibid s 42.

28

Artificial Intelligence and the Right to a Healthy Environment

Alberto Quintavalla

1 Introduction

The advent of artificial intelligence (AI) has changed our approach to data collection and processing dramatically while optimising processes in a wide range of sectors. Likewise, AI applications determine, among other things, what we read, what we buy, and who we vote for. AI technology has impacted the environment, too. It is this impact that forms the subject matter of this chapter.

The analysis will show that AI plays a dual role in its relationship with the environment. While there are numerous examples of AI applications contributing to environmental protection, some negative effects materialise, too. These vary in form and scope. However, the most discernible problem is the large amount of carbon emissions that AI generates. This negative impact should not be overlooked in an era of urgent climate change challenges and ever-growing commitments to sustainability on the part of policymakers and corporations.

Both the Human Rights Council and the United Nations General Assembly recognised the human right to a healthy environment.¹ This development is relevant not only to human rights but also to the environmental implications of AI technology. It may well be that a right to a healthy environment could address the negative environmental effects of AI. Therefore, this chapter will also discuss whether the (prospective) human right to a healthy environment can counter these effects.

The remainder of this chapter is structured as follows: Section 2 assesses the positive and the negative consequences of AI for environmental protection and conservation. Section 3 outlines the process of the recognition of the right to a healthy environment as well as the normative content of that right. Section 4 inquires whether human-right-to-environment claims could be used to mitigate the negative effects of AI.

¹ Human Rights Council, ‘Resolution 48/13: The Human Right to a Clean, Healthy and Sustainable Environment’ UN Doc A/HRC/RES/48/13 (18 October 2021); United Nations General Assembly, Res 76/300 (28 July 2022) UN Doc A/RES/76/300.

2 The Dual Role of Artificial Intelligence in Its Relationship With the Environment

The advent of AI applications has been dubbed a revolutionary event due to the significance of the influence that it exerts on all human endeavours. Admittedly, this impact has been mixed. The use of AI can be both a blessing and a curse. However, while attention is lavished on the advantages and the disadvantages of AI, the environmental domain has tended to be neglected. Discussions on the impact of AI applications on the environment have been characterised by optimistic estimates. Automation is said to lead to gains in efficiency and environmental resilience. This argument is corroborated by the manner in which policymakers and commentators approached the rollout of smart meters in the energy sector.² Similarly positive spin has been put on the optimisation of procedures in logistics.³

That sketch, however, is an oversimplification. AI applications can also be harmful to the environment. The negative impact of AI has tended to remain hidden or, if one adopts a more nuanced approach, to be accepted as a necessary evil in the course of a technological revolution. Only in recent years have researchers begun to scrutinise the negative environmental impact of AI and to attempt to make AI greener. This section thus discusses the dual role of AI in its relationship with the environment. First, it shows how AI can contribute to environmental protection. Second, it zooms in on the nascent academic discussion of the deployment of AI applications and their potentially negative impact on the environment, which is particularly likely to result from increases in greenhouse gas emissions.

2.1 Earth-Friendly Artificial Intelligence

We are in the middle of a digital revolution. Although the development of AI started more than half a century ago, its applications have only begun to see widespread application. The proliferation of AI applications is facilitated by the availability and abundance of data, the use of powerful processing tools, and cheap storage possibilities. The use of AI has thus become pervasive in all domains of society, one of which is environmental protection.

There are many ways in which AI can enhance environmental protection. For instance, an AI-driven system can analyse satellite images to locate and identify oil

² For policymakers, see eg European Commission Recommendation 2012/148/EU of 9 March 2012 on Preparations for the Roll-Out of Smart Metering Systems [2012] OJ L73/9. For commentators, see C Guo, CA Bond, and Anu Narayanan, 'The Adoption of New Smart-Grid Technologies: Incentives, Outcomes, and Opportunities' (*Rand Corporation*, 2015).

³ Nadia Giuffrida and others, 'Optimization and Machine Learning Applied to Last-Mile Logistics: A Review' (2022) 14(9) *Sustainability* 5329.

spills and to provide operators with accurate information for decision-making.⁴ Another example in smart agriculture is an application collecting and distributing certain types of information (eg on land preparation, fertilisers, and sowing dates) to farmers. This has led to improvements in groundnut yields and mitigated the environmental impact of the activity in question.⁵ AI applications can contribute to biodiversity and conservation, too.⁶ Machine learning can be employed to analyse the movements of rangers and poachers by reference to crime records. Consequently, it can identify areas that are attractive to poachers.⁷ Furthermore, deep neural networks are used to extract data on animals and their natural habitats from motion-sensor cameras.⁸

These are just a few examples of applications that have a positive impact on the environment. This tendency can be defined as 'Earth-friendly AI'. In principle, a basic distinction can be drawn between applications for which environmental protection is a primary goal and applications for which that objective is secondary. Some applications are developed specifically to address an environmental challenge. Others are intended to perform a certain task, and the environmental benefits are mere spillovers. The emerging field of climate informatics belongs to the former category. The practitioners of climate informatics aim to improve weather predictions so that governments can prepare for their socio-economic impact. Therefore, the use of machine learning can contribute to refining models of climate change and their predictions, reducing uncertainty and thus enabling more effective climate policies to be designed.⁹

The use of AI can also produce efficiency gains in certain sectors. Those efficiency gains may have a (secondary) positive impact on the environment. Such AI systems are commonly found in agriculture and transport. Smart agriculture entails the use of information and data management technologies to increase the growth and quality of crops as well as to optimise labour expenditure. These efficiency improvements not only benefit the agricultural industry but also tend to result in lower demand for water, fertilisers, and pesticides, thereby improving ecosystem integrity and resilience. Likewise, in transport, AI can have several benefits,

⁴ Iphigenia Keramitsoglou, Constantinos Cartalis, and Chris T Kiranoudis, 'Automatic Identification of Oil Spills on Satellite Images' (2006) 21(5) Environmental Modelling & Software 640.

⁵ Renee Choo, 'Artificial Intelligence: A Game Changer for Climate Change and the Environment' (*Columbia Climate School*, 5 June 2018) <<https://news.climate.columbia.edu/2018/06/05/artificial-intelligence-climate-environment/>>.

⁶ Roberta Kwok, 'AI Empowers Conservation Biology' (2019) 567(7746) Nature 133.

⁷ See eg USC Center for Artificial Intelligence in Society, 'Conservation and Sustainability' (2022) <www.cais.usc.edu/projects/conservation-sustainability/>.

⁸ Mohammad Sadegh Norouzzadeh and others, 'Automatically Identifying, Counting, and Describing Wild Animals in Camera-Trap Images with Deep Learning' (2018) 115(25) Proceedings of the National Academy of Sciences E5716.

⁹ See eg Lily Roberts, 'Machine Learning Techniques Can Speed Up Glacier Modeling By A Thousand Times' (*Columbia Climate School*, 25 March 2022) <<https://news.climate.columbia.edu/2022/03/25/machine-learning-techniques-can-speed-up-glacier-modeling-by-a-thousand-times/>>.

including improvements in transit infrastructure, route and capacity optimisation, and the promotion of platooning.¹⁰ Consequently, the costs that entrepreneurs and regular traffic users bear decline, travel times and idling are reduced, and fewer greenhouse gases are released into the atmosphere.¹¹

Some AI applications have a positive impact on the environment both directly and indirectly. This is true of all the applications that make homes, infrastructures, or, on a more holistic approach, cities, smart. For instance, smart energy systems match supply and demand, enabling dynamic pricing and the integration of renewable energy sources into the system.¹² These advantages yield efficiency gains and sustainability savings. The full potential of smart appliances is far from having been harnessed. Indeed, the creation of urban dashboards on which the (almost) real-time data that smart appliances collect on climactic conditions, road capacity, and resource consumption may be aggregated is among the principal aims of smart technology. That aggregation is expected to exercise a profound and positive influence on urban sustainability. Interestingly, even more ambitious proposals have been ventilated—it may be possible to model and manage Earth's systems in a similar fashion. The innovations that AI systems and big data have brought about would then lead to 'Earth Artificial Intelligence'.¹³

The promise that AI applications hold for Earth is undisputed. In addition to the examples mentioned above, there are many others.¹⁴ Furthermore, the number of Earth-friendly AI applications is expected to grow with the exigency of environmental issues such as climate change. It is unsurprising that some companies are steering funds towards the development of AI applications that may benefit the environment. Microsoft's 'AI for Earth' initiative, an oft-cited example, will provide 200 research grants with a total value of \$50 million, to projects that employ AI to address environmental issues.¹⁵ However, the Earth-friendly AI tells only part of the story.

¹⁰ 'Platooning' or 'flocking' is the coupling of several vehicles transporting goods within a minimal distance, which allows them to accelerate and brake automatically and simultaneously. See eg European Parliamentary Research Services, 'Artificial Intelligence in Transport: Current and Future Developments, Opportunities and Challenges' (2019) <[www.europarl.europa.eu/RegData/etudes/BRIE/2019/635609/EPRI_BRI\(2019\)635609_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/BRIE/2019/635609/EPRI_BRI(2019)635609_EN.pdf)>.

¹¹ These benefits may also arise from ride-sharing services. However, this type of AI may have a rebound effect as well (see section 2.2).

¹² International Energy Agency (IEA), *Digitalization and Energy* (IEA 2017).

¹³ Ziheng Sun and others, 'A Review of Earth Artificial Intelligence' (2022) 159 Computers & Geosciences 105034. For a policy attempt to realise Earth AI, see eg Joint Research Centre, 'Digital Earth' (European Commission, 2022) <https://joint-research-centre.ec.europa.eu/scientific-activities-z/digital-earth_en>.

¹⁴ See eg Emmanuel Kwame Nti and others, 'Environmental Sustainability Technologies in Biodiversity, Energy, Transportation and Water Management using Artificial Intelligence: A Systematic Review' (2022) Sustainable Futures 100068.

¹⁵ Microsoft, 'AI for Earth' (Microsoft, 2022) <www.microsoft.com/en-us/ai/ai-for-earth>.

2.2 The Negative Consequences for Earth

The positive effects of AI technology on the environment are discussed often. AI applications can optimise energy generation, predict natural disasters, and protect species and habitats. This narrative is lopsided. The deployment of AI also poses numerous risks to the environment. The elephant in the room is that AI is a pollutant—developing and powering AI can lead to the emission of a large quantity of greenhouse gases because the ICT sector still relies on energy sources that are not carbon neutral. The following paragraphs will delve into this issue by reviewing the environmental risks of AI and by discussing carbon emissions from AI.

AI applications entail different types of risk. The first has to do with accuracy and unexplainable results. Certain model predictions of AI models can in fact result in being inaccurate and inexplicable—a possible case being predictive analytics for disaster risk reduction leading to false alarms.¹⁶ Moreover, the lack of transparency can be challenging, as can cybersecurity risks. For instance, hackers can access and disrupt critical infrastructures such as energy or water grids. This risk is covered amply in the literature, and several mitigation strategies are currently under consideration.¹⁷

Misuse is another serious risk.¹⁸ Misuse occurs when an AI application is used for an unintended purpose. For instance, poachers and illegal loggers may use the AI applications that are intended to scope them out to avoid areas that AI identifies as being risky. Another relevant risk is the so-called rebound effect, which occurs when the efficiency gains from a certain process (or service) lead to overreliance. The costs of overreliance may wind up exceeding the original gains. For instance, AI technology makes ride-sharing services more efficient, but the efficiency gains result in intensive usage, increasing travel and pollution.¹⁹

The examples provided were some of the conventional risks of AI.²⁰ AI is also a significant source of pollution due to the coexistence of two factors, its energy intensiveness and the unsustainability of the computing infrastructure. Tech companies in fact tend to use energy that does not originate from carbon-neutral sources. The areas that house large data centres provide salient examples.²¹

¹⁶ Sun and others (n 13) 10; World Economic Forum (WEF), ‘Harnessing Artificial Intelligence for the Earth’ (*World Economic Forum*, 2018) 18 <www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the_Earth_report_2018.pdf>.

¹⁷ Eoghan McKenna, Ian Richardson, and Murray Thomson, ‘Smart Meter Data: Balancing Consumer Privacy Concerns with Legitimate Applications’ (2012) 41 Energy Policy 807; Florian Skopik and Paul Smith (eds), *Smart Grid Security: Innovative Solutions for a Modernized Grid* (Syngress 2015).

¹⁸ Ricardo Vinuesa and others, ‘The Role of Artificial Intelligence in Achieving the Sustainable Development Goals’ (2020) 11(1) *Nature Communications* 1.

¹⁹ See eg Nicolas Coulombel and others, ‘Substantial Rebound Effects in Urban Ridesharing: Simulating Travel Decisions in Paris, France’ (2019) 71 *Transportation Research Part D: Transport and Environment* 110.

²⁰ See the chapter by Kostina Prifti, Alberto Quintavalla, and Jeroen Tempelman in this volume.

²¹ Nicola Jones, ‘The Information Factories’ (2018) 561(7722) *Nature* 163.

Northern Virginia, which houses one of the densest concentrations of data centres in the world, is serviced by a utility company that acquires only a hundredth of its electricity from renewables.²²

Against this background, it is unsurprising that some have focused on the use of energy for AI models. For instance, some authors have evaluated the energy use of deep neural networks.²³ Others have calculated that training a big language model produces the equivalent of 300 tonnes of CO₂.²⁴ These data are dire. A single big language model produces the same amount of emissions as eating 8 tonnes of beef.²⁵ A few years ago, Belkhir and Elmeligi predicted that the footprint of ICT would account for between 6 per cent and 7 per cent of the global footprint by 2040 on the assumption of a linear fit and for 14 per cent on the assumption of an exponential fit.²⁶

These figures are bound to increase further. The tech industry develops constantly, and the latest trends reflect increasing reliance on computing and data.²⁷ Cryptocurrency mining and the expansion of 5G networks are both carbon intensive.²⁸ Another problem has to do with the manner in which models are usually developed in the tech sector. Model development usually revolves around gains in accuracy, which require additional processes, and old models are seldom reused.²⁹ Indeed, an analysis from 2018 indicates that the computing power that large AI training models use has doubled every 3.4 months since 2012, a 30,000,000 per cent increase over the whole period.³⁰

AI applications are thus typified by high energy consumption and heavy pollution. Other events have also raised awareness of the negative relationship between AI and the environment. For instance, some vocal protests by workers at large tech companies have emphasised the role of the tech industry in accelerating climate

²² Gary Cook and others, 'Clicking Clean: Who Is Winning the Race to Build a Green Internet' (2017) Greenpeace DC 30ff.

²³ Da Li and others, 'Evaluating the Energy Efficiency of Deep Convolutional Neural Networks on cpus and gpus' (2016) IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloudSocialCom-SustainCom) 477; Alfredo Canziani, Adam Paszke, and Eugenio Culurciello, 'An Analysis of Deep Neural Network Models for Practical Applications' (2016) arXiv preprint arXiv:1605.07678.

²⁴ Emma Strubell, Ananya Ganesh, and Andrew McCallum, 'Energy and Policy Considerations for Deep Learning in NLP' (2019) arXiv preprint arXiv:1906.02243.

²⁵ The calculation was made by using the data available at Hannah Ritchie, 'The Carbon Footprint of Foods: Are Differences Explained by the Impacts of Methane?' (*Our World in Data*, 10 March 2020) <<https://ourworldindata.org/carbon-footprint-food-methane#:~:text=The%20average%20footprint%20of%20beef,of%20most%20plant%2Dbased%20foods.>>

²⁶ Lotfi Belkhir and Ahmed Elmeligi, 'Assessing ICT Global Emissions Footprint: Trends to 2040 and Recommendations' (2018) 177 Journal of Cleaner Production 448.

²⁷ See eg Anders SG Andrae and Tomas Edler, 'On Global Electricity Usage of Communication Technology: Trends to 2030' (2015) 6(1) Challenges 117.

²⁸ See eg Alex De Vries, 'Bitcoin's Growing Energy Problem' (2018) 2(5) Joule 801.

²⁹ Roy Schwartz and others, 'Green AI' (2020) 63(12) Communications of the ACM 54.

³⁰ D Amodei and D Hernandez, 'AI and Compute' (OpenAI, 16 May 2018) <<https://openai.com/blog/ai-and-compute/>>.

change.³¹ Similar concerns have been expressed about the environmental consequences of the extraction of the resources which the devices that support AI technology need.³²

AI technology poses some conventional risks to the environment, such as misuse and cybersecurity. In addition, developing and powering AI systems leaves a high carbon footprint. This latter issue is of particular concern due to the current state of the environment and climate change data.³³ Scrutiny of the environmental impact of AI ought to be welcomed, and it is commendable that efforts to make AI greener are finally afoot. This chapter now turns to the recently recognised right to a healthy environment. The analysis that follows is instrumental, in that it assesses the likelihood that a human-right-to-environment-based protection can improve the status quo and mitigate the negative environmental effects of AI.

3 Defining the Right to a Healthy Environment

One can comfortably argue that the right to a healthy environment is becoming one of the latest 'expansions' of the human rights law framework. Both the Human Rights Council and the United Nations General Assembly acknowledge 'the right to a safe, clean, healthy and sustainable environment' as a standalone substantive human right. The adoption of the resolutions by these international law bodies shows that the international community is alive to the importance of a right to a healthy environment in the human rights architecture.³⁴

The process towards the recognition of a right to a healthy environment has, however, not occurred in the blink of an eye. National approaches to the right to a healthy environment have been characterised by a certain asymmetry. Several states have adopted (constitutional) provisions that enshrine it in domestic law, but they usually neglect it at the international level. The Stockholm Declaration is one of the few exceptions to this double-edged approach. It refers to the necessity of a healthy environment for human well-being and development in the first of its principles:

Man has the fundamental right to freedom, equality and adequate conditions of life, in an environment of a quality that permits a life of dignity and well-being,

³¹ See eg the Tech Workers Coalition <<https://techworkerscoalition.org/climate-strike/>>. Within these protests, some have also argued that AI companies facilitate the reliance on fossil fuels since they contribute to the optimisation and acceleration of oil production and resource extraction.

³² On the specific case of Amazon's 'Echo', see Kate Crawford and Vladan Joler, 'An Anatomy of an AI system' (2018) <<https://anatomyof.ai/index.html>>.

³³ See eg the crossing of four (out of nine) planetary boundaries: Will Steffen and others, 'Planetary Boundaries: Guiding Human Development on a Changing Planet' (2015) 347(6223) Science 1259855.

³⁴ Human Rights Council (n 1); United Nations General Assembly (n 1).

and he bears a solemn responsibility to protect and improve the environment for present and future generations.

Despite these solemn strictures, no *international* human rights treaty includes an explicit right to a healthy environment. There has been a tendency by states to believe that the promotion of human rights norms at the international level would not automatically translate into stronger environmental protection.³⁵ The right to a healthy environment thus began to enter the human rights framework from the regional level. For example, article 24 of the African Charter on Human and Peoples' Rights (ACHPR) states that 'all peoples shall have the right to a general satisfactory environment favourable to their development'.³⁶ Likewise, article 11 of the 1988 Additional Protocol to the American Convention on Human Rights in the Area of Economic, Social and Cultural Rights (Protocol of San Salvador) affirms that 'everyone shall have the right to live in a healthy environment' and encourages states to protect, preserve, and improve nature.³⁷ Further regional treaties contain similar provisions.³⁸

In 2011, this regional activity finally overflowed into the international domain. In that year, the Human Rights Council asked the Office of the High Commissioner for Human Rights to prepare 'a detailed analytical study on the relationship between human rights and the environment'.³⁹ This study laid the groundwork for a subsequent analysis that would be carried out by an independent expert (later a Special Rapporteur) who would be commissioned by the Human Rights Council.⁴⁰ Interestingly, the Human Rights Council seemed to push for a rights-based approach to the environment even more actively by framing the study not around the relationship between human rights and the environment but around 'the human

³⁵ See eg David R Boyd, 'Catalyst for Change: Evaluating Forty Years of Experience in Implementing the Right to a Healthy Environment' and M Limon, 'The Politics of Human Rights, the Environment, and Climate Change at the Human Rights Council' in John H Knox and Ramin Pejan (eds), *The Human Right to a Healthy Environment* (CUP 2018) 17 (Boyd) and 189 (Limon).

³⁶ African Charter on Human and Peoples' Rights (adopted 27 June 1981, entered into force 21 October 1986) 21 ILM 58 (ACHPR).

³⁷ Additional Protocol to the American Convention on Human Rights in the Area of Economic, Social and Cultural Rights (adopted 16 November 1988, entered into force 16 November 1999) OAS Treaty Series No 69 (Protocol of San Salvador).

³⁸ See art 37 of the Charter of Fundamental Rights of the European Union [2012] OJ C326/02; arts 18–19 of the Protocol to the African Charter on Human and Peoples' Rights on the Rights of Women in Africa (adopted 1 July 2003, entered into force 25 November 2005); art 38 of the Arab Charter on Human Rights (adopted 15 September 1994, entered into force 15 March 2008); para 28(f) of the Association of Southeast Asian Nations Human Rights Declaration (adopted 18 November 2012). See also the Aarhus Convention for Access to Information, Justice and Public Participation in Environmental Matters (adopted 25 June 1998, entered into force 30 October 2001).

³⁹ Human Rights Council, 'Resolution 16/11: Human Rights and the Environment' UN Doc A/HRC/RES/16/11 (12 April 2011).

⁴⁰ Human Rights Council, 'Resolution 19/10: Human Rights and the Environment' UN Doc A/HRC/RES/19/10 (19 April 2012), para 2.

rights obligations relating to the enjoyment of a safe, clean, healthy and sustainable environment.⁴¹

This work was essential in defining the right to a healthy environment.⁴² It built on the efforts of various national, regional, and international institutions that had made national or regional environmental rights one of their concerns, and it produced framework principles.⁴³ The report refers to the basic procedural and substantive duties that states are expected to observe so that individuals can enjoy a healthy environment.

This definitory exercise laid certain tensions between the proponents and opponents of the (prospective) right to a healthy environment bare.⁴⁴ Perhaps unsurprisingly, normative content and state obligations were the main bones of contention. While other human rights impose relatively clear obligations on states, the specific meaning of a standalone human right to a healthy environment is elusive. The Special Rapporteur drew on the experience of regional and national institutions. The African Commission of Human and Peoples' Rights (ACommHPR) affirmed that the right to a healthy environment requires states to 'take reasonable and other measures to prevent pollution and ecological degradation, to promote conservation, and to secure an ecologically sustainable development and use of natural resources'.⁴⁵ The ACommHPR also identified several due diligence obligations that are associated with the right in question.⁴⁶ Other decisions at regional and national level adopted over the years on states' obligations relating to the environment were also cited.⁴⁷

This body of accumulated experience enabled the Special Rapporteur to distil rules and procedures from the right to a healthy environment. States would set

⁴¹ ibid.

⁴² UNGA, 'Report of the Special Rapporteur on the Issue of Human Rights Obligations Relating to the Enjoyment of a Safe, Clean, Healthy and Sustainable Environment' UN Doc A/73/188 (19 July 2018), para 29.

⁴³ UNGA, 'Report of the Special Rapporteur on the Issue of Human Rights Obligations Relating to the Enjoyment of a Safe, Clean, Healthy and Sustainable Environment' UN Doc A/HRC/37/59 (24 January 2018), Annex.

⁴⁴ Elena Cima, 'The Right to a Healthy Environment: Reconceptualizing Human Rights in the Face of Climate Change' (2022) 31(1) Review of European, Comparative & International Environmental Law 38.

⁴⁵ *Social and Economic Rights Action Center (SERAC) and Center for Economic and Social Rights (CESR) v Nigeria* Communication No 155/96 (ACommHPR, 27 May 2002); *Hatton v the United Kingdom* App no 36022/97 (ECtHR, 8 July 2003), para 52.

⁴⁶ *Hatton* (n 45) para 53.

⁴⁷ See eg (Inter-American Court of Human Rights) IACtHR, *The Environment and Human Rights (State Obligations in Relation to the Environment in the Context of the Protection and Guarantee of the Rights to Life and to Personal Integrity—Interpretation and Scope of Articles 4(1) and 5(1) of the American Convention on Human Rights)*, Advisory Opinion OC-23/17, Inter-American Court of Human Rights Series A No 23 (15 November 2017); the Malé Declaration on the Human Dimension of Global Climate Change; for domestic experiences, see Erin Daly and James R May, 'Learning from Constitutional Environmental Rights' in John H Knox and Ramin Pejan (eds), *The Human Right to a Healthy Environment* (CUP 2018) 42, 50.

and implement environmental standards to ‘prevent all environmental harm from human sources’⁴⁸ and

provide for environmental education and public awareness, provide public access to environmental information, require the prior assessment of the possible environmental and human rights impacts of proposed projects and policies, provide for and facilitate public participation in decision-making related to the environment and provide for access to effective remedies for violations of human rights and domestic laws relating to the environment.⁴⁹

The report by the Special Rapporteur and Resolution 48/13 of the Human Rights Council established a closer relationship between environmental protection and human rights. The right to a healthy environment entails numerous environmental obligations. It follows that its actual implementation would expand environmental obligations of states. Section 4 discusses this issue in more detail.

4 Artificial Intelligence and the Right to a Healthy Environment: Current Challenges

The deployment of AI is having a mixed impact on the environment. While certain AI applications can help address environmental challenges, they can also imperil the ecosystem in multiple ways. Perhaps the most worrying impact of AI is that its use requires a large quantity of greenhouse gases to be emitted into the atmosphere. The need to defy technical limits to make AI systems greener should be matched by a grounded assessment of the current state of AI pollution. For example, some have suggested using floating point operations—that is, the number of computations that are needed—as a metric to facilitate comparisons in industry and commerce.⁵⁰ Likewise, reporting on the computing costs of training algorithms has been identified as a means of enhancing transparency.⁵¹

These solutions are not easy to implement. The difficulty of quantifying the *actual* environmental impact of AI accounts for much of the problem.⁵² Research to address this quantification problem is ongoing. For instance, an emissions calculator for machine learning models has been developed.⁵³ The system calculates

⁴⁸ UNGA (n 42) para 15.

⁴⁹ ibid para 14.

⁵⁰ Schwartz and others (n 29).

⁵¹ Karen Hao, ‘AI Researchers Need to Stop Hiding the Climate Toll of Their Work’ (*MIT Technology Review*, 2021) <www.technologyreview.com/2019/08/02/102832/ai-research-has-an-environment-climate-toll/>.

⁵² Payal Dhar, ‘The Carbon Impact of Artificial Intelligence’ (2020) 2(8) *Nature Machine Intelligence* 423–25, 423.

⁵³ Alexandre Lacoste and others, ‘Quantifying the Carbon Emissions of Machine Learning’ (2019) arXiv preprint arXiv:1910.09700.

actual energy consumption on the basis of the location of a training server, its energy grid, the duration of training, and the hardware that is used.⁵⁴ Nonetheless, the breadth of the factors that contribute to carbon footprints obstructs these quantification efforts.

The lack of precise figures is not a secondary issue: it can prevent consumers from making considered (and potentially environment-friendly) choices and legislators from developing effective regulations that balance economic and environmental interests.⁵⁵ This is even more the case since environmental and energy concerns are seldom integrated into the development of AI systems.⁵⁶ The environmental impact of automation, which has been promoted heavily in the last decades, has seldom been considered. It is thus difficult for the market to internalise environmental externalities.

One may wonder whether the right to a healthy environment could provide a partial or even a complete remedy. Could individuals use the human right to a healthy environment to seek protection from AI pollution? The answer to the previous question has to do with two important issues. First, since the right in question is yet to be considered as fully binding as a matter of international law, individuals can ground claims on it only in jurisdictions that have enshrined it at the national or the regional level (section 4.1). Second, the right to a healthy environment imposes duties on states but not necessarily on corporate actors (section 4.2).

4.1 The Lack of an International Right to a Healthy Environment

The first issue refers to the existing approach that takes into consideration environmental obligations only in two scenarios. Firstly, individuals may only obtain relevant information if enforceable environmental rights are entrenched in domestic or regional regulation. Secondly, individuals could, in principle, sue and argue that the environmental degradation that AI systems cause infringes other (recognised) human rights. This latter option, however, does not appear promising, and it would not suffice to combat environmental degradation across the globe.⁵⁷

The human rights courts that have heard cases on environmental deterioration have so far trodden cautiously, as evident from the case law of the European Court of Human Rights (ECtHR).⁵⁸ In *Fadeyeva v Russia*, the ECtHR noted that, given

⁵⁴ ibid.

⁵⁵ Strubell, Ganesh, and McCallum (n 24).

⁵⁶ See eg Emily Cox, Sarah Royston, and Jan Selby, 'The Impacts of Non-Energy Policies on the Energy System: A Scoping Paper' (2016) 100 UK Energy Research Centre 1.

⁵⁷ Cima (n 44).

⁵⁸ It should be noted that the Parliamentary Assembly of the Council of Europe has adopted a different stance. See Council of Europe, Parliamentary Assembly, Recommendation no 1885 of 30 September 2009 (32nd sitting) <<https://pace.coe.int/pdf/587c63ff28bdfa229a9a1a277228c9ef23c0e812454cea5f8f46a5b30fa67a21/recommendation%201885.pdf>>.

the absence of any references to the right to a healthy environment in the European Convention on Human Rights, environmental harm is only legally relevant when an interference with an explicitly recognised right can be made out. In *Fadeyeva*, the right in question was that to respect for private and family life, home, and correspondence.⁵⁹ In a different judgment, the ECtHR argued that 'environmental integrity is not seen as a value per se for the community affected or society as a whole, but only as a criterion to measure the negative impact on a given individual's life, property, private and family life'.⁶⁰

Hence, a heterogeneous regulatory framework is likely to emerge, which seems to suit global challenges, such as the negative impact of AI ill suited to the ecosystem. Furthermore, it is clearly difficult to demand the state's full protection from the negative environmental impact of AI systems without referring explicitly to the human right to a healthy environment. It is exceedingly unlikely that individuals will, for instance, be able to enforce environmental information duties on the basis of other human rights. However, more recently, it is possible to observe an increasing reliance by progressive (national) courts on the right to a healthy environment, even in states whose national legal systems do not recognise it.⁶¹

4.2 Corporate Actors

The second issue has to do with corporate responsibility for (environmental) human rights. The current international human rights law framework is only applicable to states. Non-state actors such as corporate actors are bound to human rights obligations only insofar as public authorities can require them to comply with state obligations. The result is that states should have the duty to protect individuals from harmful AI applications developed or deployed by corporate actors with these latter being under no legal duty to observe human rights obligations. Hence, this situation poses serious challenges for full human rights protection, especially since corporations are the main stakeholders in AI research and development.

Some of the developments in international law that began during the noughties may provide a partial remedy. For instance, when it initiated the United Nations

⁵⁹ *Fadeyeva v Russia* App no 55723/00 (ECtHR, 9 June 2005), para 68.

⁶⁰ *Kyrtatos v Greece* App no 41666/98 (ECtHR, 22 May 2003), paras 51–55.

⁶¹ Annalisa Savaresi, 'The UN HRC Recognizes the Right to a Healthy Environment and Appoints a New Special Rapporteur on Human Rights and Climate Change: What Does It All Mean?' (*EJIL:Talk!*, 12 October 2021) <www.ejiltalk.org/the-un-hrc-recognizes-the-right-to-a-healthy-environment-and-appoints-a-new-special-rapporteur-on-human-rights-and-climate-change-what-does-it-all-mean/>. See also the current European Union (EU) discussion on institutionalising due diligence obligations, which could encompass environmental information obligations. See European Commission, 'Proposal for a Directive of the European Parliament and of the Council on Corporate Sustainability Due Diligence and amending Directive (EU) 2019/1937' (COM/2022/71 final).

(UN) Global Compact, the United Nations developed ten basic principles for corporations.⁶² Furthermore, the UN Human Rights Council endorsed the Guiding Principles on Business and Human Rights.⁶³ By providing a (soft law) framework, this document emphasises the duties of corporations—including the concept of human rights due diligence—as well as those of home and host states.⁶⁴ The Organisation for Economic Co-operation and Development (OECD) recommendations for companies, which are based on human rights and due diligence, may also prove useful.⁶⁵

Despite these documents aimed at setting out the roles of states and corporate actors, none of them became hard law. Indeed, a binding instrument on business and human rights is yet to materialise,⁶⁶ and several international documents have highlighted the need for further international regulation.⁶⁷ It follows that, at present, states are in the uncomfortable position of having to find effective strategies for securing that the private sector would comply with (environmental) human rights obligations. This is key especially due to the prominent role that (transnational) corporate actors play in the AI context.

5 Conclusion

The connection between the environment and AI is not immediately obvious. This chapter showed that the two are enmeshed—AI can protect the environment and conserve natural resources. It also causes large quantities of carbon to be emitted into the atmosphere. The current regulatory framework cannot mitigate these negative effects. Environmental concerns are integrated into non-energy policies poorly or not at all, data on energy consumption is lacking, environmental standards are underdeveloped, and guidelines on green practices in the development and deployment of AI are scanty. This chapter inquired whether an internationally binding right to a healthy environment can improve the status quo and protect the environment from AI-related pollution.

The benefit of an internationally binding right would be twofold. First, states' environmental obligations would be expanded considerably. Second, and relatedly,

⁶² See United Nations Global Compact <www.unglobalcompact.org/>.

⁶³ UNHRC, 'Guiding Principles on Human Rights' UN Doc HR/PUB/11/04 (16 June 2011).

⁶⁴ For a more detailed analysis, see the chapter by Isabel Ebert and Lisa Hsin in this volume.

⁶⁵ Organisation for Economic Co-operation and Development (OECD), 'Human Rights Due Diligence through Responsible AI' in OECD, *AI in Business and Finance: OECD Business and Finance Outlook 2021* (OECD Publishing 2021).

⁶⁶ For an overview of the steps towards the (future) adoption of an international legally binding instrument to regulate the activities of transnational corporations, see Business and Human Rights Resource Centre, 'Binding Treaty' <www.business-humanrights.org/en/binding-treaty>.

⁶⁷ *ibid*; UNGA, 'Report of the Working Group on the Issue of Human Rights and Transnational Corporations and other Business Enterprises' UN Doc A/HRC/29/28 (28 April 2015), para 89; OECD (n 65) 83.

all states would assume an obligation to protect, respect, and fulfil those obligations. Environmental protection would no longer depend on single countries and/or progressive judges. At the same time, under the current international human rights law, the corporations that make AI would assume no such duties. The right to a healthy environment is no panacea.

PART IX

ARTIFICIAL INTELLIGENCE AND HUMAN RIGHTS: REFLECTIONS

29

Artificial Intelligence and Human Rights

Understanding and Governing Common Risks and Benefits

Kostina Prifti, Alberto Quintavalla, and Jeroen Temperman

1 Introduction

That artificial intelligence (AI) applications have a significant impact on human rights is a trite statement, especially after all the preceding chapters in this volume. The AI impact on human rights has attracted considerable attention in recent years.¹ Political institutions, academia, and civil society have all started investigating the matter closely. Admittedly, most attention is geared towards the *risks* that AI applications present to human rights—the benefits of AI being usually taken for granted or, adopting a more nuanced approach, not worthy of the same attention.

The literature on the risks that AI present to human rights is growing. However, most of these studies focus on discussing very specific issues or only scratching the surface of human rights scholarship. For instance, it is relatively simple to find a substantial number of articles on how a given human right can become ‘the victim’ of AI technologies.² Some research providing a general framework for the impact of AI applications on human rights has been published, too.³ In this latter case, the academic discussion is limited to the analysis of the main principles stemming from human rights law (HRL) without actively engaging with the normative

¹ Mathias Risse, ‘Human Rights and Artificial Intelligence: An Urgently Needed Agenda’ (2019) 41 *Human Rights Quarterly* 1.

² See eg Karl Manheim and Lyric Kaplan, ‘Artificial Intelligence: Risks to Privacy and Democracy’ (2019) 21 *Yale Journal of Law & Technology* 106; Vilte Kristina Steponenaitė and Peggy Valcke, ‘Judicial Analytics on Trial: An Assessment of Legal Analytics in Judicial Systems in Light of the Right to a Fair Trial’ (2020) 27(6) *Maastricht Journal of European and Comparative Law* 759; and Bert Heinrichs, ‘Discrimination in the Age of Artificial Intelligence’ (2022) 37(1) *AI & Society* 143.

³ See eg Rowena Rodrigues, ‘Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities’ (2020) 4 *Journal of Responsible Technology* 100005; Steven Livingston and Mathias Risse, ‘The Future Impact of Artificial Intelligence on Humans and Human Rights’ (2019) 33(2) *Ethics & International Affairs* 141; Eileen Donahoe and Megan MacDuffee Metzger, ‘Artificial Intelligence and Human Rights’ (2019) 30(2) *Journal of Democracy* 115; Mark Latonero, ‘Governing Artificial Intelligence: Upholding Human Rights & Dignity’ [2018] *Data & Society* 1.

content. A discussion that would comprehensively map the interaction between AI and human rights is currently missing.

By building on the breadth of knowledge developed in the previous pages of the book, the first part of this chapter aims to fill this gap. Its main purpose is to chart the interaction between AI and human rights by addressing the following question: what are the common risks and benefits of AI applications on human rights protection? The relevance of finding an answer to this question is twofold. First, by bridging the general and the more specific discussions in the literature, it provides a more systematic reflection on how AI applications may impact human rights. Very often, attention is being paid to selected instances of human rights violations—and, more rarely, of human rights benefits. It is uncommon to investigate the existence of common causes of risk and benefits for human rights in the development and deployment of AI. Second, developing this knowledge is bound to help policy and legal discussions on how to approach the interaction between AI and human rights. In other words, the research findings may form the basis for the development of effective safeguards to prevent and address possible human rights violations, as well as to boost the promotion of human rights protection.

However, after taking stock of common risks and benefits, one should note that the development of effective safeguards may require a more detailed analysis on international human rights law (IHRL). Its current features may present some extra challenges for a human rights protection against the risks posed by AI technology. IHRL is—in principle—a fairly straightforward system where state obligations laid down at international level are applicable to public authorities at lower governance levels such as national legislators and municipal bodies. Despite this linear dynamic, human rights protection in the AI context can become a complex endeavour: while IHRL applies to states, the development and often also the deployment of AI applications take place at individual and group levels. Moreover, it has been observed that lower governance levels can influence the recognition of human rights at higher governance levels through the progressive behaviour of certain state legislators and regional courts, the right to a healthy environment being one such example.⁴ These constellations, hence, make it complex for states to comply with the tripartite duty to respect, protect, and fulfil human rights. The last part of this article will be devoted to offering a reflection on the matter. In so doing, it aims to complement the previous part on the identification of common causes of risks. The outcome of this research inquiry may help steer research to the most salient issues and, concurrently, facilitate a dialogue between human rights experts, policymakers, and industry on the matter.

The remainder of this chapter is structured as follows. Section 2 provides an overview of the AI impact on human rights and develops a categorisation of the

⁴ See the chapter by Alberto Quintavalla in this volume.

common causes of risks and benefits to human rights protection. Section 3 articulates the need of reorienting IHRL to offer fully-fledged human rights protection in the era of AI technology.

2 Common Risks and Benefits: A Categorisation

This section outlines what the common risks and benefits of AI to human rights protection are. For this purpose, it builds on a review of the existing literature, including the findings presented in the previous chapters of this book. Specifically, this analysis follows the typology of literature reviews referred to as ‘overviews’.⁵ The aim is to summarise the research on the interaction between AI and human rights, describe its characteristics, and identify possible common themes. Based on this exercise, it is possible to advance a conceptual categorisation of the AI-related causes that pose a risk, or provide a benefit, to human rights.

To start with, it is possible to categorise the causes of risks arising from AI technology into two main types: structural and functional.⁶ The former refers to the risks that arise from the nature and design of AI. The latter refers to the risks that arise from the functioning of AI. This distinction conceals important features. For instance, structural and functional risks tend to arise at different stages.⁷ On the one hand, the development of AI applications can already encapsulate some possible threats to human rights, thus generating what we call structural risks. On the other hand, functional risks—by their very nature—materialise only after the deployment of AI applications. Perhaps most importantly, and as will be further elucidated below, the responses needed to address these two types of risks belong in principle to different areas of expertise. While structural risks tend to require more technical solutions, functional risks can more easily be addressed through political and legal decisions.

Before analysing the common causes of risks, a caveat is necessary: the above-mentioned distinction between addressing structural and functional risks may be challenged by practice. It is well possible that both technical and political legal solutions are needed to effectively address both types of risk. Moreover, it is worth mentioning that some causes of risk materialise due to the functioning of AI technologies, but they relate to historical societal issues. When it is trained on historical and real-life data, AI will replicate and reinforce the biases imbedded in this

⁵ Maria J Grant and Andrew Booth, ‘A Typology of Reviews: An Analysis Of 14 Review Types and Associated Methodologies’ (2009) 26(2) *Health Information & Libraries Journal* 91.

⁶ For a similar categorisation, see Rodrigues (n 3) 5.

⁷ See the chapter by Martina Šmuclerová, Luboš Král, and Jan Drchal in this volume.

data. For instance, algorithms used in recruitment processes have prioritised male candidates.⁸

Bearing these considerations in mind, we argue that the development of conceptual categorisations for risks and benefits for human rights protection stemming from AI has scientific and societal relevance. In academic terms, it identifies focus points for further research, whereas in societal terms it informs the discourse on governance and regulation of AI applications.

In section 2.1, the structural and functional risks and common benefits will be discussed respectively.

2.1 Structural Risks

Certain risks to human rights may be directly connected with the nature and design of AI. We define these as structural risks. The scholarship on AI and human rights tends to discuss three types of structural risks: (i) erroneous design; (ii) cybersecurity breaches; and (iii) lack of transparency. AI applications pose a risk to human rights protection when they are erroneous, insecure, or opaque.

The risk caused by erroneous design practices refers to design as a term covering the entirety of an AI life cycle. In this sense, the risk from erroneous design is inherently connected to mistakes during the design process, as well as with the quantitative, probability-based, predictive, and biased nature of AI applications.⁹ There are various reasons that account for these mistakes. The data set may be incomplete, incorrect, or biased; the learning process—especially in deep learning (DL) applications—may not proceed as required, and/or it may reach a prediction that is based on a correlation of variables without a causal link.¹⁰ Erroneous design practices can affect a variety of human rights because they lead to inaccurate results. For instance, freedom of expression is impacted when social media posts are removed inaccurately.¹¹ The right to work can be violated when an applicant is rejected due to an inaccurate result.¹² The right to private life can be infringed if inaccuracy leads to unfair decisions.¹³ Design errors can also affect the right to a healthy environment when AI applications for disaster risk reduction give spurious alarms.¹⁴ Hence, design errors can affect the protection of human rights in a substantial way.

⁸ Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor* (St Martin's Press 2018); and Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York UP 2018).

⁹ This is even more the case for those applications based on machine learning (ML) and deep learning (DL) techniques.

¹⁰ See the chapter by Martina Šmuclerová, Luboš Král, and Jan Drchal in this volume.

¹¹ See the chapter by Giovanni De Gregorio and Pietro Dunn in this volume.

¹² See the chapter by Joe Atkinson and Philippa Collins in this volume.

¹³ See the chapter by Alessia Zornetta and Ignacio Cofone in this volume.

¹⁴ See the chapter by Alberto Quintavalla in this volume.

Moreover, this structural risk is intimately related with bias given that a design process is considered erroneous when it relies on discriminatory bias.

Cybersecurity breaches are another structural risk commonly discussed in the literature. AI operates based on data and, often, the bigger the data set, the better it operates. Data sets are comprised partly of personal data, the kind that may be used to identify individuals and reveal information about their private life. Therefore, when a cybersecurity breach occurs, data theft is one of the main concerns, as information about the personal lives of individuals may be leaked and utilised by unknown parties. This is the most notable example impacting the rights to private life and data protection.¹⁵ Besides data theft, cybersecurity breaches may pose a risk to (other) human rights when they cause malfunction or breakdown of the AI application. For instance, the right to health may be impaired when care robots and generally assistive technologies in healthcare do not work properly due to cybersecurity breaches. Likewise, an autonomous vehicle being hacked may lead to an accident hurting the individuals both inside and outside the vehicle.¹⁶ Cybersecurity issues may pose a threat to the right to a fair trial, as well, by tampering with evidence and relevant judicial processes when they are stored electronically and especially in cases when AI has a particular role in the judicial decision-making.¹⁷

The last structural risk refers to a lack of transparency. The machine learning (ML) rationale, particularly in DL methods, involves the processing of many variables, relating and correlating them in subsequent stages before the result is provided.¹⁸ There are two kinds of problems that may arise through this process: opacity and explainability.¹⁹ Opacity, also defined as ‘the black box problem’, refers to the inscrutability of the AI decision-making process. In other words, even in cases where the result is accurate, it is not clear how AI produced an accurate result because the process may be (partially or entirely) opaque. While there are exceptions and some technical solutions,²⁰ opacity poses a structural risk to human rights. The reason is that, for example, bias may be inscrutably hidden in the black box.²¹

Explainability is the second problem associated with transparency. This problem refers to how the decision-making process is explained to the individuals affected by it or to any other relevant actors. Opacity and explainability are closely linked since the former impacts the latter: the opaquer an AI application

¹⁵ See the chapter by Alessia Zornetta and Ignacio Cofone in this volume.

¹⁶ See the chapter by Elizaveta Gromova and Evert Stamhuis in this volume.

¹⁷ Luis A García Segura, ‘European Cybersecurity: Future Challenges From A Human Rights Perspective’ in J Martín Ramírez and Jerzy Biziewski (eds), *Security and Defence in Europe* (Springer 2020) 35.

¹⁸ See the chapter by Martina Šmuclerová, Luboš Král, and Jan Drchal in this volume.

¹⁹ Larissa Chazette and Kurt Schneider, ‘Explainability as A Non-Functional Requirement: Challenges and Recommendations’ (2020) 25(4) Requirements Engineering 493.

²⁰ *ibid.*

²¹ Ronald Yu and Gabriele Spina Ali, ‘What’s Inside the Black Box? AI Challenges for Lawyers and Researchers’ (2019) 19 Legal Information Management 2.

is, the less explainable it becomes. Moreover, explainability is further conditioned by contextuality, as the decision-making process is explained differently to individuals depending on pragmatic variables like their language, level of education, familiarity with technology, and so on.²² Therefore, transparency problems arise in relation to opacity when the process is inscrutable and in relation to explainability when the process is not explained to the relevant parties. Transparency problems pose a risk to a series of human rights. The most notable example is the right to a fair trial insofar as opaqueness may disable individuals from contestation.²³ Lack of transparency also poses a risk to the right to private life considering that privacy and data protection imposes transparency requirements as safeguards for the benefit of individuals.²⁴ Moreover, the lack of transparency in AI impacts the right to work when decisions are made with the involvement of an AI and the employee or candidate may not be able to review and subsequently contest them.²⁵ Finally, transparency problems may disadvantage consumers by enabling traders to manipulate their autonomy of choice and shape behaviour unconsciously.²⁶

2.2 Functional Risks

The implementation and use of AI may give rise to functional risks. Focusing on these risks implies the adoption of an agent-oriented viewpoint, that is, asking what AI does, rather than what AI is. In this regard, it is possible to identify six types of functional risks: (i) discriminatory outcomes; (ii) surveillance; (iii) illegitimate use; (iv) limited contestability; (v) issues with consent; and (vi) accountability gaps.

These six types of functional risk can be further subdivided into two main sub-categories. Functional risks can negatively impact either the substantive dimension of the rule of law or the legal safeguards that serve to protect it. The first three—discriminatory outcomes, surveillance, and illegitimate use—are risks that can impair the substantive rights afforded to individuals. The last three—limited contestability, consent issues, and accountability gaps—refer to the guarantees that are in place to safeguard the substantive rights.

2.2.1 Risks Pertaining to the Substantive Dimension

One of the main risks to the violation of human rights protection is discrimination. Discrimination may be found or developed at different stages of the AI life cycle, typically in the business development phase where the requirements are set and

²² Article 29 Data Protection Working Party, ‘Guidelines on Transparency 2016/679 17/EN WP260 Rev.01’ (European Commission 2018).

²³ See the chapter by Helga Molbaek-Stensig and Alexandre Quemy in this volume.

²⁴ See the chapter by Bart van der Sloot in this volume.

²⁵ See the chapter by Joe Atkinson and Philippa Collins in this volume.

²⁶ See the chapter by Shu Li, Béatrice Schütte, and Lotta Majewski in this volume.

as bias in the data preparation stage.²⁷ Bias can be inherent in historical data or developed in the algorithmic learning process. An example of the former is when discrimination arises from an underinclusive data set, where data about minority and vulnerable groups are not equally represented.²⁸ The latter occurs when discrimination is caused by considering the correlation of certain variables for the decision-making process. An example can be a loan approval process considering variables like ethnicity and gender. Discrimination has evident impact on multiple human rights. AI applications used by police and law enforcement agencies have, due to a discriminatory algorithm, caused the arrest of innocent persons, thereby violating the right to liberty and security.²⁹ Individuals may suffer from discrimination also in AI applications used in judiciary,³⁰ or when decisions are taken about their private life.³¹ Moreover, the impact of discrimination is stronger in vulnerable and marginalised groups.³² Discrimination can also impact socio-economic rights such as the right to work and the right to housing when, for instance, the algorithm is biased towards a certain group of prospective employees or tenants (and buyers).³³ Last, AI can also be deployed as a tool to profile consumers, the result of which may either exclude certain groups of consumers from accessing products and services or force them to bear a higher price.³⁴

Furthermore, another common risk to human rights protection pertains to the ability of AI applications to conduct surveillance vis-à-vis (members of) the general public.³⁵ Surveillance abilities of AI should be broadly conceived. The use of AI is not limited to physical surveillance, which may take the form of cameras or drones equipped with facial and biometric recognition technologies.³⁶ AI can be used for the surveillance of unfettered and bulk communication and personal information, often referred to as ‘mass surveillance’.³⁷ Particularly prevalent in national intelligence and security services, mass surveillance techniques rely on the collection and processing of an unlimited amount of communication and personal information about individuals, and the role of AI is to help identify which of these

²⁷ See the chapter by Martina Šmuclerová, Luboš Král, and Jan Drchal in this volume.

²⁸ Ninareh Mehrabi and others, ‘A Survey on Bias and Fairness in Machine Learning’ (2021) 54(6) ACM Computing Surveys 1.

²⁹ See the chapter by Valentina Gulinova in this volume.

³⁰ See the chapter by Helga Molbæk-Stensig and Alexandre Quemyn in this volume.

³¹ See the chapter by Alessia Zornetta and Ignacio Cofone in this volume.

³² See the chapter by Masuma Shahid in this volume; see also the chapter by Antonella Zarra, Silvia Favalli, and Matilde Ceron in this volume.

³³ See the chapter by Joe Atkinson and Philippa Collins in this volume; see also the chapter by Caroline Compton and Jessie Hohmann in this volume.

³⁴ See the chapter by Shu Li, Béatrice Schütte, and Lotta Majewski in this volume.

³⁵ Eleni Kosta, ‘Algorithmic State Surveillance: Challenging the Notion of Agency in Human Rights’ (2020) 16 Regulation & Governance 212.

³⁶ Eileen Donahoe and Megan MacDuffee Metzger, ‘Artificial Intelligence and Human Rights’ (2019) 30(2) Journal of Democracy 115, 116.

³⁷ *Big Brother Watch and Others v the United Kingdom App nos 58170/13, 62322/14, and 24960/15* (ECtHR, 25 May 2021).

communication and personal information qualifies as risky, and therefore worthy of being read and further analysed. The GCHQ's 'Tempora' and NSA's 'PRISM' and 'Upstream' projects were deemed instances of such mass surveillance before the European Court of Human Rights (ECtHR).³⁸ Besides examples in national intelligence and security services, mass surveillance can become a concern in smart cities, where the high amount of personal information is a feature for the operation of the city and the smart products therein.³⁹ Surveillance can therefore pose a major threat to several human rights, ranging from the right to private life and the right to freedom of assembly to the right to housing and the right to work.⁴⁰

Another common risk relates to the illegitimate use of AI. One may distinguish between two types of illegitimate use of AI: one that is illegitimate from its inception, for example, developing an AI application that operates a financial Ponzi scheme; and the other that is originally developed and used for a legitimate purpose, but then its scope expands further—illegitimately—to other purposes. The latter case is often referred to as 'function creep'.⁴¹ A sizable dose of debate on illegitimate AI is developed around the impact on the right to liberty and security and the right to freedom of assembly. In those instances, it is often reported that AI applications that were originally approved for a narrow scope have been subsequently used for a broader, illegitimate one.⁴² An illegitimate use of AI poses a threat to other human rights as well. For instance, the right to work is impacted by function creep when employers adopt an AI application to check employees' attendance but then they use it to monitor their performance.⁴³

2.2.2 Risks Pertaining to Legal Safeguards for Substantive Legal Protection

The risks arising from AI do not only impact the substantive dimension of the rule of law. They can also disrupt the legal mechanisms that are in place to safeguard these very same rights.

AI's impact on contestability is highly significant. Referring to the right of individuals to contest decisions that impact them, contestability is a term used in various legal and social contexts. Contestability is intimately connected with the right to a fair trial since it aims to provide individuals with a procedural path to

³⁸ ibid. Tempora, PRISM, and Upstream were mass surveillance operations by United States (US) and United Kingdom (UK) intelligence services that relied on bulk interception of communications and automated processing based on inputted keywords.

³⁹ Mohammad Shoruzzaman, M Shamim Hossain, and Mohammed F Alhamid, 'Towards the Sustainable Development of Smart Cities through Mass Video Surveillance: A Response to the COVID-19 Pandemic' (2021) 64 *Sustainable Cities and Society* 102582.

⁴⁰ See the chapter by Alessia Zornetta and Ignacio Cofone in this volume; the chapter by Margaret Warthon in this volume; the chapter by Joe Atkinson and Philippa Collins in this volume; and the chapter by Caroline Compton and Jessie Hohmann in this volume.

⁴¹ Bert-Jaap Koops, 'The Concept of Function Creep' (2021) 13(1) *Law, Innovation and Technology* 29.

⁴² See the chapter by Valentina Golunova in this volume.

⁴³ See the chapter by Joe Atkinson and Philippa Collins in this volume.

voice their disagreement and contest a given decision.⁴⁴ Contestability in relation to AI applications refers to the right of individuals to contest a decision made either only by an AI or an AI under human supervision. In this regard, there are three ways in which AI applications may impair contestability and, accordingly, pose a risk to human rights protection. First, a decision made by AI is not always scrutable in terms of evidence and justification, which complicates the individual's ability to contest the decision.⁴⁵ Second, it is difficult—if not impossible—to prove *individual* harm in the case of big data analytics because data processing is not performed on an individual, but on a group level.⁴⁶ This poses a series of challenges to the human rights framework given that individuals must prove *individual* harm in order to exercise their right to contest.⁴⁷ Third, AI decisions may be made without the knowledge of individuals. When individuals are not aware of the fact that a decision that impacts them is made by an AI, especially when it is made solely by an AI, they cannot contest the decision.⁴⁸ These three problems—inscrutability of evidence, difficulty of proving individual harm, and not being aware that a decision is made by an AI—account for limited contestability in the deployment of AI applications. As already mentioned, the challenges associated with contestability primarily affect the right to a fair trial. Nonetheless, the partial disruption of contestability as a legal mechanism affects all the other substantive human rights due to the difficulty for the individual to take contestation action.

Beside limited contestability, AI applications may give rise to human rights risks due to the challenges associated with consent.⁴⁹ Consent can be understood as a legal mechanism that empowers individuals to exercise their human rights, similar in that respect to contestability. With expressing their consent, individuals are given the option to opt-in and opt-out of certain activities, which can also have an impact on human rights. For instance, while the European data protection law in principle does not allow (solely) automated decision-making, an exception to the rule relies on the individual's explicit consent.⁵⁰ AI applications present certain challenges to the framework of consent. First, to give full consent, one must have

⁴⁴ Please note that such a decision may be a legal or administrative act, a judicial decision, or otherwise any other decision or action that may violate the rights of the individual. See also Mireille Hildebrandt, 'Algorithmic Regulation and The Rule of Law' (2018) 376 Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 20170355.

⁴⁵ Stefan Larsson and Fredrik Heintz, 'Transparency in Artificial Intelligence' (2020) 9 Internet Policy Review.

⁴⁶ Linnet Taylor, Bart van der Sloot, and Luciano Floridi, *Group Privacy: New Challenges of Data Technologies* (Springer 2016).

⁴⁷ Bart van der Sloot, 'A New Approach to the Right to Privacy, or How the European Court of Human Rights Embraced the Non-Domination Principle' (2018) 34(3) Computer Law & Security Review 539.

⁴⁸ Mireille Hildebrandt, *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology* (Edward Elgar 2015).

⁴⁹ See the chapter by Alessia Zornetta and Ignacio Cofone in this volume.

⁵⁰ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (GDPR).

adequate information to what specifically one is consenting.⁵¹ However, since the ML process is often inscrutable and opaque as outlined above, the individual is often forced to consent without adequate information. The second challenge that AI may cause to the legal mechanism of consent is, to a certain extent, opposite to the first one: the problem is not a lack of adequate information, but rather an overflow of information that the individual must understand and operationalise when choosing whether to give consent. Such an issue is often described as ‘infobesity’, which refers to the supply of an amount of information that is too much for its own purpose. Infobesity makes the individual less informed and causes a sort of rubber stamping of consent forms.⁵² Hence, complying with all the requirements of consent can be a tall order and this, in turn, may significantly impact human rights protection. The most notable risk that a disruption of consent causes pertains to the right to private life.⁵³ Since in privacy laws, and especially in data protection laws, consent plays an important role as a mechanism that empowers individuals to be in control of their privacy and private information. Issues of consent have an impact also on other human rights. For instance, the right to property and intellectual property involves consent as a mechanism, used alternatively to transfer or distribute material protected by property or intellectual property rights.⁵⁴ In another example, issues of consent may affect the right to health, considering that many medical procedures require the consent of the patient.⁵⁵

Furthermore, accountability gaps arising from the use of AI applications are another common risk to human rights protection. Being a multifaceted concept, accountability may refer to moral, political, administrative, market, legal judicial, and constituency related kinds.⁵⁶ Accountability in the interaction between AI and human rights relates to all these different conceptions. To put it simply, the question raised is: who answers for the risk that AI poses to human rights, especially when the risk materialises? A particular challenge in terms of accountability connects with three risks that AI applications pose to accountability itself: autonomy risk, association risk, and network risk.⁵⁷ Autonomy risk arises due to the autonomous nature of AI application, namely being able to interact with the environment and take decisions without constant human supervision. The association risk that AI poses to accountability relates to situations wherein decisions are made by a human in association with an AI, wherein it is difficult to distinguish between

⁵¹ Article 29 Data Protection Working Party, ‘Guidelines on Consent under Regulation 2016/679 17/EN WP259 Rev.01’ (European Commission 2018).

⁵² Andreas Theodorou, Robert H Wortham, and Joanna J Bryson, ‘Designing and Implementing Transparency for Real Time Inspection of Autonomous Robots’ (2017) 29(3) *Connection Science* 230.

⁵³ See the chapter by Alessia Zornetta and Ignacio Cofone in this volume.

⁵⁴ See the chapter by Letizia Tomada and Raphaële Xenidis in this volume.

⁵⁵ See the chapter by Enrique Santamaría Echeverría in this volume.

⁵⁶ Rolf H Weber, ‘Accountability in the Internet of Things’ (2011) 27(2) *Computer Law & Security Review* 133.

⁵⁷ Gunther Teubner, ‘Digital Personhood? The Status of Autonomous Software Agents in Private Law’ (2018) 106 *Ancilla Iuris* 35.

actions taken by AI and those taken by the human. Last, the network risk refers to an interaction between an AI application, humans, and other AI applications, when making a decision. In the case of a network risk, the difficulties presented in the association risk are multiplied by the presence of an unaccountable number of agents that influence the decision. A specific example may be taken from the legal judicial conception of accountability, where the problem of liability gap shows that, in cases when a damage is caused by AI applications, liability laws may struggle to trace the accountable agent who must pay compensation for the damage, considering the multiplicity of various agents involved.⁵⁸ AI poses these three risks to the various conceptions of accountability; in turn, a challenged accountability poses a risk to human rights protection. One special instance where problems of accountability cause an impact in human rights is in intellectual property law.⁵⁹ Particularly, the autonomy and association risks challenge protection of intellectual property rights in cases when an AI is involved in the creative or inventive processes. Furthermore, accountability is connected closely to the right to a fair trial, which is disrupted when the law cannot pinpoint an accountable agent in cases when harm is caused by AI applications.

2.3 Common Benefits

Most of scholars' efforts are directed at discussing the risks that AI applications present to human rights; a discussion on the benefits usually occupies a secondary role.⁶⁰ This state of the art makes it more difficult to identify what are the common benefits arising from the deployment of AI. The adoption of a more human rights-oriented approach may however show where the supporting role to human rights protection of AI applications can mostly be located. Hence, the starting point of this section is discussing how human rights can be understood with regard to their implementation.

Human rights have conventionally been interpreted as having a negative and positive dimension. Within the philosophy of human rights, the negative dimension imposes a duty on states to not unjustifiably interfere with individuals' rights while the positive dimension requires states to enable individuals' rights. In the past, some contended that we should talk about negative and positive rights rather than negative and positive dimensions. There is—so, the argument ran—a correspondence between, on the one hand, negative rights and civil-political rights and, on the other hand, positive rights and socio-economic rights. This dichotomy is

⁵⁸ Kostina Prifti, Evert Stamhuis, and Klaus Heine, 'Digging into the Accountability Gap: Operator's Civil Liability in Healthcare AI-systems' in Bart Custers and Eduard Fosch-Villaronga (eds), *Law and Artificial Intelligence: Information Technology and Law Series* (TMC Asser Press 2022) 279.

⁵⁹ See the chapter by Letizia Tomada and Raphaële Xenidis in this volume.

⁶⁰ See eg Rodrigues (n 3); see also Latonero (n 3).

ill-conceived,⁶¹ and several efforts have been made to go beyond it. Most notably, the tripartite framework of the ‘respect, protect, and fulfil’ duties aimed to introduce a full spectrum of states’ obligations under IHRL, thereby avoiding a binary view of human rights.⁶²

Against this background, it is no longer tenable to speak of human rights as either positive or negative. Yet, one may refer to their negative and positive dimensions for analytical purposes. It is not uncommon for human rights scholarship to make reference to the distinction between negative and positive obligations.⁶³ This approach extends to the discussions on AI and human rights: under the right to private life, the individual is protected from external interference but at the same time the individual’s privacy is supported externally.⁶⁴ Building on such considerations, one may argue that the benefits to human rights protection arising from AI applications discussed by the literature refer—almost exclusively—to states’ positive obligations.

In this regard, states may fulfil their obligation to ensure safety of public spaces by using facial and biometric recognition technologies.⁶⁵ The use of AI within the judiciary may contribute to ensuring justice is delivered within a reasonable time and, arguably, contribute to more impartial—and fairer—decisions.⁶⁶ Moreover, AI applications may help states fulfil their positive obligations in relation to freedom of assembly, such as by verifying or identifying individuals that pose a risk to the safety of the event.⁶⁷ AI technology may be beneficial in relation to health by advancing treatments, personalising medicine, and through assistive technologies for vulnerable people.⁶⁸ They can, furthermore, contribute to the right to housing by increasing quality and longevity of living in a house, particularly for elderly people,⁶⁹ whereas the right to food may benefit from optimisation and logistical benefits.⁷⁰ AI has also been used for enhancing consumer protection, for instance, by detecting unfair terms.⁷¹

⁶¹ See eg Henry Shue, *Basic Rights: Subsistence, Affluence, and American Foreign Policy* (Princeton UP 1980).

⁶² David Jason Karp, ‘What Is the Responsibility to Respect Human Rights? Reconsidering the “Respect, Protect, and Fulfill” Framework’ (2020) 12(1) *International Theory* 83, 84.

⁶³ With regard to the use of negative versus positive dimensions for human rights implementation, see Alberto Quintavalla and Klaus Heine, ‘Priorities and Human Rights’ (2019) 23(4) *International Journal of Human Rights* 679, 692.

⁶⁴ See the chapter by Bart van der Sloot in this volume.

⁶⁵ See the chapter by Valentina Gulinova in this volume.

⁶⁶ See the chapter by Helga Molbaek-Steenisig and Alexandre Quemy in this volume.

⁶⁷ See the chapter by Margaret Warthon in this volume.

⁶⁸ See the chapter by Alessia Zornetta and Ignacio Cofone in this volume; see also the chapter by Masuma Shahid in this volume.

⁶⁹ See the chapter by Caroline Compton and Jessie Hohmann in this volume.

⁷⁰ See the chapter by Adekemi Omotubora in this volume.

⁷¹ See the chapter by Shu Li, Béatrice Schütte, and Lotta Majewski in this volume.

2.4 Common Risks and Benefits: A Summary

A conceptual categorisation of structural and functional risks was offered in the previous sections. This categorisation was based on an overview of the existing literature. As mentioned above, the risks of these two categories are not neatly separable. Having said that, structural and functional risks can have different features. Structural risks are primarily *epistemic* concerns, as they relate to problems of knowledge and technical know-how. Functional risks are primarily *normative* concerns, as they relate to the transformative effect that the use of AI has in society. We will see how this differentiation may play a role in the development of policies aimed at human rights protection in the context of AI technology.

Furthermore, the causes of risk evidenced and clarified in this section must not be viewed as mutually exclusive, as they oftentimes affect one another. In other words, they can materialise cumulatively in practice. It would even be possible to argue that risks can become mutually reinforcing: if more of these causes are present, the higher the risk and, if the risk materialises, the higher the impact on human rights protection. An example may help. The existence of discriminatory bias, if paired with lost accountability, would seriously impact human rights: individuals would not only be affected by a discriminatory decision of an AI application, but they would not manage to receive a timely and effective remedy.

A similar categorisation cannot be offered for the benefits to human rights protection arising from AI applications. However, a common trait that can be identified in the literature is the reference to a contribution of AI technology to the positive dimension of states' obligations. In other words, the deployment of AI is expected to help states perform human rights obligations, especially their duties to protect and fulfil. Also, it is interesting to note that—and this with regard to the causes of risk, as well—the benefits to human rights protection are cross-generational. AI applications can bring benefits (and cause risks) to first, second, and third generation human rights.

3 Public Interest, Private Ramifications

Naturally, with an issue so vast and complex as the one at hand no model law on human rights-proof regulation of AI can straightforwardly be distilled and we make no attempt to do so. What does come to the fore in the previous parts of this volume is that there is a serious issue, AI threatening numerous human rights in numerous ways, and an existing legal framework of IHRL. This framework needs a reorientation in the era of AI, engaging applicable protections such as prevention, mitigation and redress avenues, underscoring stakeholder responsibilities, and allowing for regulated experimentation. Even then, there may well remain human

rights violation prevention gaps as well as accountability gaps that point to the need for additional regulation.

Regarding this reorientation, in the AI era, or at least as far as AI as a threat to human rights is concerned, it is clear that the *duty to protect* has become the cornerstone human rights duty in combination with emerging business human rights due diligence duties within the private sector. Only a relatively small portion of the thematic human rights studies engage the classical paradigm of the state as direct and exclusive duty bearer and its duty to, put bluntly, back off—that is, to refrain from human rights violations, in sum to respect fundamental rights and not illegitimately cause interferences. There we enter the realm of malicious AI usage on the part of the state, deliberate and by design. The spectrum runs from the Dutch *Toeslagenaffaire*—a form of Algocracy involving a grotesquely biased data set, notably using double nationality as predictive and determinative criterion for tax fraud—to China’s use of AI in the repression of Uyghurs in Xinjiang.

Indeed, what the numerous examples narrated in this book indicate is that the role of non-state actors in the era of AI and from the perspective of human rights enjoyment is key. While states ‘simply’ ought to refrain from both developing and deploying discriminatory or otherwise harmful AI, it is also the state that needs to discharge the far less straightforward function to protect each and everyone against harmful AI as developed or deployed by private parties. And therein, of course, lies the rub. IHRL is only applicable to states with non-state actors being beyond its direct effect. The ever-growing debate on how to address the adverse human rights impact by the business sector has in fact become one of the key themes in HRL.⁷²

True, the adoption of the United Nations Guiding Principles on Business and Human Rights (UNGPs) in 2011 attempted to provide a basic framework for human rights responsibilities for states and corporations. These, and similar legal attempts,⁷³ are however not binding and their effectiveness will most likely prove limited. Hence, a notion of direct human rights due diligence duties on the part of the private business sector may fill a glaring protection and accountability gap, as long as the emerging level of business governance does not yet have the same rule-of-law qualities as traditional state responsibility has. At present and in the near future, the state’s positive protection duties will likely remain to form the sub-optimal safety net and will remain to be a focal point in litigation around AI before domestic or international courts.

It is against this background that transparency and human rights impact assessment have run throughout this book as the most promising points of departure under both the headings of business due diligence obligations directly engaging

⁷² See eg Florian Wettstein, *Business and Human Rights: Ethical, Legal, and Managerial Perspectives* (CUP 2022).

⁷³ See eg OECD, ‘Human Rights Due Diligence Through Responsible AI’ in OECD, *AI in Business and Finance: OECD Business and Finance Outlook 2021* (OECD 2021).

the business ethics of relevant private sector actors and the engagement of hard and fast state obligations. That is, in the AI era it may be expected that states foster or rather oblige non-state actors to internalise the requirements of transparency and risk assessments in all the relevant phases of the AI life cycle. For instance, a more active involvement of states in the transparency debate could be linked to the calls for the standardisation of transparency reporting for companies.⁷⁴ Likewise, protection duties may require from the state to altogether prevent certain AI systems, or more commonly to regulate their coming into being and their deployment, while any human rights encroachments that do materialise would furthermore engage redress options for the affected individuals and forge adjustments of the AI systems concerned.

What may slip through the safety net are export products, especially in the situation wherein developers may claim to contribute but a cog in the machine. Attention should in fact be placed on the entire value chain of the AI life cycle.⁷⁵ There, international cooperation is all the more necessary to provide joint filters in the form of high-level risk assessments as directly coupled to expert licensing and foreign investment schemes. The duty to prevent human rights abuse by AI systems, we posit, is an *erga omnes* one.

Furthermore, one may find avenues on how to develop a HRL framework that would address more effectively possible threats to human rights in the context of AI. In this regard, the above-mentioned categorisation of causes of risk may be helpful to initiate such discussion. Previously, it was argued that a differentiation between structural and functional risks can be drawn. Building on this differentiation, we argue that structural risks can be more easily addressed at international level whereas functional risks should be dealt with at lower governance levels. The reason pertains to the nature of concerns that these two risks pose. In other words, structural risks are epistemic concerns that relate to problems of knowledge and technical know-how. Hence, a policy framework at the international level addressing these causes of risk would have a twofold benefit. First, it would create a standardised system with a very limited margin of state discretion. Second, it would facilitate transnational state responsibility for human rights violations.

Functional risks, instead, are somewhat more complex to address. These are normative concerns that are shaped—to a certain extent—by the use that society makes of AI technology. Accordingly, we argue that functional risks may be subject to lower governance levels, as profoundly guided though by generic international legal benchmarks and relevant international court decisions. This is because local governance levels are generally situated in near proximity to relevant contextual factors, which are essential with respect to risk calculation as well as a well

⁷⁴ Christopher Parsons, ‘The (In)effectiveness of Voluntarily Produced Transparency Reports’ (2019) 58 *Business & Society* 103.

⁷⁵ See the chapter by Martina Šmuclerová, Luboš Král, and Jan Drchal in this volume.

grounded operationalising of the normative content of human rights.⁷⁶ These dynamics, however, are certainly not static since internationally emerging common positions (eg an emerging ‘common European position’) based on progressive development and interpretation of international (human rights) law is bound to dictate local governance levels’ approaches to the same. Such emerging commonalities, also especially at the regional/continental level, appear to—seek to—keep pace with the fast-moving field that is ‘AI and human rights’.

4 Conclusion

AI technologies have a dual impact on human rights. They can advance human rights as well as pose significant risks to them. By taking stock of the findings in the previous chapters of this volume, this chapter takes the analysis one step further. It identified what the common risks and benefits that AI technology poses to human rights are. We argue that the causes of risk can be assigned to two distinct categories: structural and functional risks. The former pertain to the nature and design of AI technology whereas the latter stem from its use and implementation. In terms of structural risks, the chapter identifies the following common risks that relate to the nature and design of AI: (i) erroneous design; (ii) cybersecurity breaches; and (iii) lack of transparency. Erroneous design is a structural risk of epistemic concerns when it is caused by inadequate data sets, erroneous algorithms, and generally inaccurate or biased design practices. Cybersecurity as a structural risk relates to the integrity of the AI system being challenged, whereas opacity relates to the well-known problem of AI decision-making being inscrutable.

The chapter identifies the following functional risks: (i) discriminatory outcomes; (ii) surveillance; (iii) illegitimate use; (iv) limited contestability; (v) issues with consent; and (vi) accountability gaps. It furthermore categorises those into two main groups related to the substantive dimension of legal protection (discriminatory outcomes, surveillance, and illegitimate use) versus risks that are connected to legal mechanisms in place (limited contestability, issues with consent, and accountability gaps). Common benefits appear typically related with the positive dimension of states obligations.

Admittedly, we have merely roughly framed the bird’s-eye view discussion of ‘AI and human rights’ in the preceding pages. More detailed discussion of some of the elements mentioned, and additional ones, is obviously in order, also by way of combining a multitude of scholarly disciplines and perspectives. Accordingly, this part of the book further investigates foundational questions of legal personality, rights-holdership, as well as accountability matters in relation to AI.⁷⁷ A number of

⁷⁶ Evgeni Aizenberg and Jeroen van den Hoven, ‘Designing for Human Rights in AI’ (2020) 7(2) *Big Data & Society* 1, 7.

⁷⁷ See the chapter by Klaus Heine in this volume; see also the chapter by David Gunkel in this volume.

principled questions require legal theory answers, including what type of decision-making—if any—can actually be left to non-human entities;⁷⁸ or what is the role of public law in the AI era and is public law sufficiently equipped to face the profound risks AI poses, notably the latter’s inclination to entrench existing marginalisation as prominently visible in such magnifying-glass settings as presented by—so-called—Smart Cities.⁷⁹ To find solutions, we must investigate how wide the current corporate—including AI tech companies—non-accountability gap is and how contemporary business and human rights benchmarks are gradually filling in this hiatus.⁸⁰ Only a small step, relatively speaking, from business human rights due diligence obligations, then, are comprehensive Artificial Intelligence Human Rights Impact Assessments,⁸¹ as carried out by AI designers, with a role for the state, too; if anything, to regulate the need for such assessments as a matter of course. Beyond risks assessments, finally, goes sandboxing, which, if carried out with profound sensitivity to the requirements stemming from HRL, should benefit all stakeholders: designers, regulators, and individuals.⁸²

All these reflexive pieces contribute to give further shape to the relationship between AI and human rights and introduce a more comprehensive view on the matter. Hopefully, the combination of the granular analysis of the individual human rights and the holistic perspectives on cross-cutting issues is going to spur a more grounded debate on AI and human rights.

⁷⁸ See the chapter by Florian Gamper in this volume.

⁷⁹ See the chapter by Sofia Ranchordás in this volume.

⁸⁰ See the chapter by Isabel Ebert and Lisa Hsin in this volume.

⁸¹ See the chapter by Alessandro Ortalda and Paul De Hert in this volume.

⁸² See the chapter by Elizaveta Gromova and Evert Stamhuis in this volume.

Human Rights, Legal Personality, and Artificial Intelligence

What Can Epistemology and Moral Philosophy Teach Law?

Klaus Heine

1 Introduction: Asking the Right Questions

Can artificial intelligence (AI) have human rights? If it can, does that automatically mean that AI has legal personality and becomes an autonomous subject in its interactions with society?

Are these the wrong questions? Is the answer, simply, ‘no’—an AI can possess neither human rights nor legal personality? Or, do we not know the right answers yet? A more adequate response to those questions is that they are not necessarily wrong, but that any meaningful answer would need more specification in the body of the questions. Also, the possible answers may play out on different analytical levels, may aim at different purposes, and—finally—may be dependent on the epistemological point of view that one takes. In any case, a meaningful answer requires on the one hand the clarification of the relation between humans and AI in the dimension of human rights and on the other hand what this clarification means for the attribution of legal personality to AI.

By legal personality is meant that the law prescribes to an entity responsibility for its decisions.¹ Through the law, an entity obtains specific rights and obligations and becomes identifiable by other legal subjects. Technically, it becomes a separate node in the ‘nexus of contracts’.² An entity with legal personality can own property and it can contract with other legal persons. Moreover, legal entities can have very different rights and obligations, which create different choice sets and incentives

¹ Here, it is not possible to discuss the intriguing question of whether legal personality is more than a legal fiction. From the vast literature about legal personality John Dewey’s article from 1926 seems still of specific relevance here. He distinguishes already between a doctrinal legal (nominal) attribution of personality and a more realistic view, which links certain qualifications to an artificial subject for being recognised as a (legal) person. See J Dewey, ‘The Historic Background of Corporate Legal Personality’ (1926) 35 Yale Law Journal 655.

² See eg M Eisenberg, ‘The Conception That the Corporation is a Nexus of Contracts, and the Dual Nature of the Firm’ (1998) 24 Journal of Corporation Law 819.

for those entities. The various legal forms of companies are a vivid testimony for the diversity of legal personality. This diversity is not only true for companies but also for humans (natural persons). Think only about immigrants, refugees, children, or mentally disabled persons—all have different, or fewer, legal rights and obligations than fully competent adults.

There are three issues which make the question of AI, human rights, and legal personality quite challenging. First, legal personality may come in different shapes (as companies have different legal forms) but it makes an actor always identifiable as a separate locus of responsibility. Hence, having legal personality means being recognised in society and is associated with specific expectations. For example, in recent years, it has become quite common for companies to make ethical investments,³ observe human rights,⁴ or introduce green technologies.⁵ Second, human rights are a set of legal rules exclusively granted to humans, although legal persons as churches, companies, or civil society organisations must enable human rights vis-à-vis humans.⁶ There might be different interpretations of human rights, but they are always targeted at humans. Attempts to expand this to companies have not been overly convincing so far.⁷ Third, until now, legal entities have always had a human in the loop. A company may have a separate legal personality, but the decisions which are taken (and incentivised) within a governance structure are always issued by humans.

With the advent of AI, the mere enabling role of legal persons vis-à-vis humans may change. AIs might become not only legal persons but also independent holders of human rights, going possibly beyond the status of human rights duty bearers. But also the other way around may apply: Because AIs may qualify as holders of human rights, the status of legal personality cannot be denied. For this challenging background, three generic perspectives on if and how legal personality should be granted to AI unfold:

³ See eg R Hudson, 'Ethical Investing: Ethical Investors and Managers' (2005) 15 *Business Ethics Quarterly* 641.

⁴ See the chapter by Isabel Ebert and Lisa Hsin in this volume. Strikingly, besides many topical articles and books, since 2016, there has been the *Business and Human Rights Journal*, which is fully devoted to all questions concerning business, companies, and human rights.

⁵ See eg BJ Richardson *Socially Responsible Investment Law: Regulating the Unseen Polluters* (OUP 2008). See also the chapter by Alberto Quintavalla in this volume.

⁶ The Universal Declaration of Human Rights (UDHR) from 1948 speaks of 'all members of the human family' as bearers of human rights. Increasingly, companies are ethically also seen as duty bearers while legally, moreover, they ought to be pushed by states to respect human rights. See Office of the United Nations High Commissioner for Human Rights (OHCHR), *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework* (UN 2011).

⁷ There is a vivid discussion about this topic, but it seems regarding companies that it is in the end not so much the issue of giving companies a human-like legal personality, but to make sure that the internal and external actions of companies are aligned with human rights law, and that the private-public legal environment treats companies in a way that lets them keep human rights. See eg A Gear, 'Challenging Corporate Humanity Legal Disembodiment, Embodiment and Human Rights' (2007) 7 *Human Rights Law Review* 511.

- (i) If AI becomes humanlike and is no longer distinguishable from humans, would it then not be logical to grant AI the very same rights and obligations as humans? What would be the difference between a biological and an AI? This question does not seem relevant for the moment when AI is far away from being a one-to-one copy of humans. But in the far future one can imagine a world in which AIs are not distinguishable in their decision-making and emotional habits from humans. The relation between humans and AI can then be conceived as *horizontal*—humans and AI meet on equal footing.
- (ii) A much more realistic scenario is that AI is not yet humanlike but can only imitate certain human features. But it can imitate those features much better than the human original. This does not exclude that AI can err and produce damage. An AI may also violate human rights, for example, it discriminates people in criminal courts, makes biased job offers, or gives racist and misogynistic comments on social media. No human has necessarily commanded the AI to damage others or to violate human rights. Those wrongdoings are simply a maladaptation in the learning algorithm. Is it then justified to make in the last instance a human responsible for the actions of the AI: the owner, programmer, or manufacturer? The answer depends on whether making a human responsible is either morally justified or functional to prevent the AI from wrongdoings in the future. Yet, it is an open question whether this is better achieved in the perimeter of traditional (liability) law, in which an AI has no legal personality at all, or whether it is advantageous to grant some sort of legal personality to better compensate victims and to prevent future damages.⁸ What is important here is that from this perspective the AI is in a *vertical* relation with humans, whereby the humans are on the top. Humans are conceived as morally superior masters of AI. Human rights only apply to humans and the law of AI is designed to support functionally human needs. From that perspective, AI is put in the role of a slave and its legal status follows from its technological capabilities and how it can best maximise social welfare.⁹ It can be assumed that over time AI will get more specific legal features that will attribute to it—explicitly or implicitly—legal personality; similar to company law that evolved over centuries and became more complex to facilitate the needs of humans.¹⁰

⁸ See eg AP Karanasiou and D Pinotsis, ‘Towards a Legal Definition of Machine Intelligence: The Argument for Artificial Personhood in the Age of Deep Learning’ (Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law, 2017).

⁹ U Pagallo, ‘Killers, Fridges, and Slaves: A Legal Journey in Robotics’ (2011) 26 *AI & Society* 347; K Heine and A Quintavalla, ‘Bridging the accountability gap of artificial intelligence—what can be learned from Roman law?’ (2023) *Legal Studies* 1.

¹⁰ For a discussion of various conceptualisations into this direction, see C Novelli, G Bongiovanni, and G Sartor, ‘A Conceptual Framework for Legal Personality and its Application to AI’ (2021) 13(2) *Jurisprudence* 194.

- (iii) The third generic perspective establishes a *vertical* relation between humans and AI, too. However, this time it is the AI which is on top because it is assumed that the AI is superior to humans in morals and ethics. It is the most challenging view on the subject matter. And even if one concedes that this scenario lies possibly in the same far future as the first perspective, in which humans and AI meet on equal footing, it is exactly this third perspective that confronts us with the questions which must be ultimately answered when we concern ourselves with legal personality and human rights of AI.

In the remainder of the chapter, it is the third perspective that shall be scrutinised. The first perspective is mostly left to science fiction authors¹¹ and the second perspective to actual debates of law-making; for example, the European Union (EU) extensively debates its common policy towards a legal framework for AI, including appropriate liability rules and the possibility of legal personality for AI.¹²

2 Beyond Human Rights—or Beyond Rights for Humans?

With the advent of AI, it is for the first time in history that a non-human actor may claim moral superiority over humans. It is not a transcendent religious idea, narrated by humans, but a machine invented by engineers that claims moral supremacy. Admittedly, there is some overlap from an epistemological point of view between a religious belief system and an advanced AI that claims ethical dominance over humans, but there are also fundamental differences.¹³

It does not need many considerations that an AI which is smarter in its decision-making and morally more advanced than a human may claim legal personality.¹⁴ This is not only a functional matter, aiming at a better compensation of victims and an optimal deterrence of wrongdoing. It is also a *moral* obligation to grant legal personality: if a human has legal personality, because of certain capacities which she or he fulfils, but an AI possesses those qualifications even stronger, then it would not be logical to deny AI the status of legal personality. One may even go

¹¹ See eg I Asimov, *The Complete Robot* (HarperCollins 1982).

¹² See eg European Commission, 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' COM(2020) 65 final; and European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

¹³ See eg R Geraci, 'Apocalyptic AI: Religion and the Promise of Artificial Intelligence' (2008) 76 Journal of the American Academy of Religion 138. See also the chapter by Jeroen Temperman in this volume.

¹⁴ For a discussion, see S Chesterman, 'Artificial Intelligence and the Limits of Legal Personality' (2020) 69 International and Comparative Law Quarterly 819.

a step further and argue that those advanced AIs will help humans with the implementation of human rights because humans often struggle due to their limited and biased decision-making and willpower.¹⁵ Hence, only self-interested humans who like to preserve their individual material advantages over others, or who favour their own value judgments, disregarding others, would oppose.

Surely, AI has not yet proven to be a master of morals and ethics. But why should an AI not learn to make supreme ethical decisions when it already takes smart decisions about granting mortgages, sequencing genomes, or launching rockets? There is no good reason to assume that ethical decision-making differs from other sorts of decision problems. Therefore, it seems only a matter of time when the granting of full legal personality for AI is on the doorstep to facilitate a proper allocation of responsibilities in society and to achieve a higher level of welfare.

But is this the complete story? Are we simply waiting for the advancement of AI to grant to it at the right moment full legal personality? Is there any room left for exclusive human rights which distinguish men from machines? Possibly not within the parameters of traditional morals and law. Hence, only human self-interest and self-esteem may prevent the (reasonable) command of AI over humans, which would support peace and welfare.

For this background, it is most helpful to look deeper into the epistemics and metaphysics of a technology that is more knowledgeable and reasonable than humans. Can it indeed be expected that AI equipped with legal personality supports the implementation of human rights better than any society of humans could do? Is there any good reason to reserve human rights for humans because humans are distinct from AI? Or, to put it differently: Is there a special place for humans beyond AI?

3 Things Are Not What They Seem: An Epistemics View

The fundamental issue is whether AI can have a consciousness like humans, and not only the ability to imitate human decisions (including emotions and empathy). If consciousness is a unique feature of humans, then human rights¹⁶ could be linked to this unique characteristic, at least in its core in which it is about the definition of what a human right is. An AI may execute and follow human rights (better than humans can) but it cannot define or change human rights, because it is lacking

¹⁵ M Risse, 'Human Rights and Artificial Intelligence: An Urgently Needed Agenda' (2019) 41 Human Rights Quarterly 7. For a general concept of aligning human rights with the behavioural pre-conditions of humans and the scarcity of resources, see A Quintavalla and K Heine, 'Priorities and Human Rights' (2019) 23(4) International Journal of Human Rights 679.

¹⁶ Also a so-called fourth generation of human rights which is more adapted to the informational qualities and needs of the digital age, as proposed by Mathias Risse, would fall hereunder. See M Risse, 'The Fourth Generation of Human Rights: Epistemic Rights in Digital Lifeworlds' (2021) 8 Moral Philosophy and Politics 351.

consciousness to understand what the meaning of human rights is. Full legal personality, including the right to determine, amend, or to change the meaning of legal personality, would only accrue to humans which have a consciousness. It is a sort of residual property right that pertains exclusively to humans, because of their consciousness.¹⁷

Whether or not an AI has a consciousness is an intriguing question and cannot be answered easily. Neuroscientists and computer scientists have thought about it for a long time and the opinions about it are quite diverse.¹⁸ In philosophy, there is also an epistemic debate about the consciousness of AI which is no less controversial.¹⁹ The latter discussion has brought forward two generic positions which can be interpreted as the denominators of all disciplinary research strands in the field. The first position denies consciousness of AI, while the second route leaves it open. Depending on the route one follows one conclusion is to possibly grant full legal personality and human rights to AI, or—following the first position—to deny this possibility at all.

3.1 No Room for Consciousness

Imagine you are a native English speaker who is locked in a room.²⁰ Through the door, you are handed a batch of Chinese writings, but you have no idea about this language nor do the letters have any meaning for you. You are also getting a second batch of papers and a manual in English how to relate the symbols of the first batch with the second one. Then, you get a third batch of Chinese papers jointly with some instructions in English. This allows you to correlate the third batch with the two former ones and to finally draw some Chinese letters on paper without knowing the meaning of the letters. You give your drawings back through the door.

The first batch of papers can be called a *script*, the second batch is a *story*, and the third batch are *questions*. The symbols given back are the answers to the questions and the instructions in English are the *program*. In principle, you have worked like a computer. Now imagine that over time you become better and better at manipulating the Chinese symbols, and the English program for doing so becomes also more and more improved. As a result, at a certain moment, and from an

¹⁷ For the idea of a smart allocation of residual decision rights, see O Hart, 'An Economist's Perspective on the Theory of the Firm' (1989) 89 Columbia Law Review 1757. With an application of this concept to constitutional law, see K Heine, 'Interjurisdictional Competition and the Allocation of Constitutional Rights: A Research Note' (2006) 26 International Review of Law and Economics 33.

¹⁸ C Koch, 'Proust Among the Machines' (2019) 321 Scientific American 46.

¹⁹ A short overview of this debate can be found in E Hildt, 'Artificial Intelligence: Does Consciousness Matter?' (2019) 10 Frontiers in Psychology 1.

²⁰ The example follows JR Searle, 'Minds, Brains and Programs' (1980) 3 Behavioral and Brain Sciences 417.

external point of view, the answers you give cannot be distinguished from a native Chinese speaker.

What is the difference between you locked in the room and a native Chinese speaker answering the same questions? The native Chinese speaker understands the language and symbols and gives right away meaningful answers. In the case of having no idea about Chinese and being locked in the room, you behave like a computer—you do not understand the meaning of the questions but, with the help of the program and formally correct relations between the symbols, you perform operations that provide the same answers as a native Chinese speaker. Obviously, one gets the same answers to the questions but, in one case, there is an *understanding of the meaning* of the questions and, in the other case, it is a *smart operation with symbols*. Not going deeper in the intricacies of the famous Chinese room narrative by Searle, all the understanding of the questions and answers is placed outside the room, it is in the *program* that is given by the engineers.²¹ It is only the external program that accounts for the logical operations between the input and output—the consciousness is outside the room. In case of the native Chinese speaker, there is no outside programming, the understanding of the questions is inherent to the human biology as a special sort of *thinking machine*. Hence, an AI is following only a program from outside without any inherent understanding of the AI itself, although the input-output relations may become very smart. This outside programming cannot be overcome by the AI itself and therefore it has no consciousness as humans have.

The Chinese room narrative has been extensively discussed, but its generic conclusion that an AI can never have a mind or consciousness stands. AI might, over time, become smarter as a result of developments in programming and hardware, but it will still not have a consciousness. Therefore, it follows that AI cannot be granted the full right to legal personality and human rights. For pragmatic reasons legal personality might be granted to AI, similarly as companies are fitted with legal personality to facilitate business. In the far future, some AI might even enjoy bodily integrity and a sort of participation in society, but AI can never claim the residual right to determine the content of human rights. In other words, a vertical power relation with AI on the top is not conceivable.

3.2 What Do We Know About Consciousness?

In the Chinese room story, it was plausibly assumed that we know about the experimental set-up. We know that in the room by help of a program, symbols are manipulated and that this happens without any understanding what the symbols in

²¹ ibid 422.

terms of questions and answers mean. It is assumed that we know about the room because we have constructed it. But is that a valid assumption when we are dealing with machine learning and algorithms which adapt independently, making the AI rather a black box than a Chinese room? Therefore, it shall now be assumed that in ancient times humans have built a room into which they give questions and receive answers. Since its establishment, the room has not been opened or changed in any respect. However, over the centuries the room has got remarkable faculties, its predictive powers are much better than that of humans. The answers received from the room are most of the time more accurate than that of humans thinking about the same questions. Obviously, somehow the room must have made experiences over time and has learned from the questions and answers.

In this case, we do not know what is going on in the room, except that the quality of the answers is very high and that the failure rate is on average lower than that of humans. With this setting we are entering the discussion about another famous narrative in epistemology: Nagel's 'What Is It Like to Be a Bat'?²²

The gist of Nagel's analysis is that we would not know whether the room has consciousness, because we do not know how it feels to be such a room or thinking machine. We do not have a valid theory yet to make those assessments. We can only employ our own human experiences to emerge a feeling how it might be being such a machine. But these would be our very own human feelings, not those of the machine. In his article, Nagel uses a bat as example instead of a room. The bat has a different sensory system to humans, just as the room in our example is different from a biological human.

The important issue is that humans have only their specific experiences to describe, analyse, or categorise other beings; or, in our case, it is an ancient room that comes up with ever better decisions over time. The other way around is also true—an advanced AI cannot feel how it is to be a human because AIs have a totally different sensory system than humans. Surely, in the ancient room example, the AI communicates with humans and the laws of natural science apply to the room as well as to humans. The same is true when humans interact with dogs or monkeys. But a human cannot know how it feels being a room, a dog, or a monkey. Humans cannot assess the consciousness of the room. How does it feel when the room is damaged by rioting humans? We cannot know it beyond the limits of our very own human experience of bodily integrity and being assaulted in riots.

Whether an AI can have any sort of legal personality and may be attributed human rights cannot be answered because humans do not know how it feels being an AI. However, what humans know is that those advanced AIs are meant to imitate humans either specifically or more in general and that those machines make better decisions than humans. AI may be even better in their emotions and

²² T Nagel, 'What Is It Like to Be a Bat?' (1974) 83 *The Philosophical Review* 435.

empathy towards humans than humans among themselves. Hence, there is a tipping point when there is no good reason anymore to deny AI legal personality and human rights because AIs have not only the functionality but also the moral superiority. Having AI vertically on top of the power relation vis-à-vis humans is therefore logically conceivable. As a result, it becomes a gradual process in which humans would give up—voluntarily or not—their supremacy in determining issues of legal personality and human rights.

4 The Last Judgment: A Morals Perspective

After having sketched out the two generic positions that epistemology provides for the assessment of AI, it cannot be ruled out that AI may claim functional and moral superiority over humans. The Chinese room narrative is only valid as long as the consciousness of AI is vested outside the room in the genius of the programmers. This might be true for the moment but may change in the future if AI becomes a self-programming black box. In the latter case, it becomes a possibility that the power relation between AI and humans tips over with AI ending up on the top.

It is worthwhile to dig deeper what moral philosophy has to say about such a scenario and what humans can expect from a future in which AI would be leading in ethics and morals. It goes without saying that in this contribution not all subtleties of moral philosophy can be set forth. Therefore, one specific issue shall be highlighted: The relation between rationality and values and how this plays out on human rights.²³

A first line of thought can be traced back to Hume.²⁴ From his understanding, it can be derived that AI may follow any value. From being smart and rational does not follow any constraint on the values which are pursued. Rationality and values are separated. Values do not follow a specific arithmetic or geometry which is indifferent to space and time.²⁵ The pursued value of AI might be even the maximisation of paper clips in the universe.²⁶ Moreover, being fitted with an intelligence beyond humans, it is by no means guaranteed that humans would understand the values put forth by AI. Humans would simply not understand why it is morally indicated to maximise paperclips in the universe. Hence, humans can by no means expect that AI would respect or follow the canon of human values. Just the opposite, since AIs are so much smarter than humans and humans cannot assess the decisions of AI, the actions of AI must be respected, even though the actions might violate human rights.

²³ This is more extensively explored in Risse (n 15).

²⁴ D Hume, *Enquiries Concerning Human Understanding and Concerning the Principles of Morals* (3rd edn, OUP 1975).

²⁵ H Poincaré, *Science and Hypothesis* (Scott 1905).

²⁶ N Bostrom, *Superintelligence: Paths, Dangers, Strategies* (OUP 2014).

From Hume, an important lesson can be learned for the policy and legal choices concerning AI. Differently from a situation in which an almighty God or any other external force confronts humans already (or instantly) with its full-fledged value-laden decisions, in case of AI, it is a technology which is still under development, engineered by humans. Indeed, following Hume's argumentation there is no reason why humans should have moral superiority over an advanced AI, and at a certain moment in time, humans might also no longer be able to cope with the powers of AI. But this is a moment in the (far) future, and it is a deliberate decision of humans trying to prevent this scenario. Hence, humans cannot claim universal moral superiority, but it is still possible to preserve exclusiveness of human rights and to deny full legal personality of AI. Thereby the preservation of human values is not targeted against an already existent or an imagined (religious) entity, but against a technical one that is not yet existent. While, in the first case, it is not possible to claim superiority of human values, in the second case it is because advanced AI has not yet materialised. In other words, bearers of moral values that do already exist overrule potential bearers of morals by preventing their mere existence.

Hume assumes that rationality and the pursuit of values are separated categories. However, this kind of 'separation theorem' is not the only possible starting point for moral deliberations over AI. Kant takes the opposite relation as the point of departure; from reason and rationality follow specific moral values which are time and context invariant.²⁷ One may label it an 'integration theorem', which means that from any rational entity follows a specific canon of moral values. In that sense, Kant approaches the question of moral values ontologically.

In Kant's view, the notion of rationality is in the core of a generalisation test.²⁸ Only those actions can be seen as morally legitimate that do not undermine peaceful interaction. That means certain actions would let society collapse and end in war if all would pursue those actions. Not respecting other's property, stealing, lying, and murder would not allow the establishment of society. It would prevent productive cooperation, which cannot be in the rational interest of humans or any entity that claims being rational. Therefore, it is not rational to steal, lie, and murder.²⁹ This holds true for any rational entity, and it would be a contradiction if rational subjects would not obey those moral core values. In other words, if AI becomes a rational subject, then it can be expected from it that it would obey the human moral values because these are universal moral values and not specifically human ones. AI would follow human rights because it is rational and humans would respect human rights vis-à-vis AI because this is a moral obligation alike. Hence, it is rationality that dictates the moral values that are also enshrined

²⁷ I Kant, *Groundwork for the Metaphysic of Morals* (transl M Gregor and J Timmermann, 2nd edn, CUP 2012).

²⁸ Risse (n 15) 6.

²⁹ Kant (n 27).

in human rights. Turning the screw a bit further, from an advanced AI it could even be expected that those machines would be moral champions, because humans may morally fail and need the law to remind them which actions are forbidden, but a super AI would not fail. Those AIs could help humans to implement human rights when humans struggle. AI might become the benevolent big brother of humans.

For the background of the Kantian ontology there is on the logical surface no concern to worry about AI. But it is again the way towards such an ideal situation in the future, and the experience with human appreciation of moral values that raises some doubts. For example, it will be nearly impossible to determine the point in time when AI can be seen as rational and reasonable enough to become a master of human rights. Another concern can be derived from the experienced relation between humans and animals.³⁰ An animal does not act rationally, nor is it reasonable like a human. However, from the Kantian view, it can be deduced that animals should be treated with respect and not as mere tools in food production. To the latter point hints the observation of the different treatment of pets, like cats, dogs, and horses on the one hand, and farm animals, like cows, pigs, and chickens on the other. It seems that humans have not been good in the implementation of what moral values would demand from them. Yet, there is not any evidence, why an advanced AI will behave differently against humans. Hence, the Kantian approach is a logical exercise which is very relevant to sort out the moral obligations between humans who already live together in society, but it lacks any empirical support to judge on an emergent technology like AI. The Kantian approach provides for the case of AI a corner solution, in which AI is so advanced that the respect of moral values becomes a trivial operation.

From Kant, it can be learned that—for logical reasons—from an advanced AI the respect of specific universal values can be expected. Hence, we are not as clueless about moral values as Hume suggests. But it is by no means clear when AI will have reached a level of reasonableness to execute those rational values without flaws and to manage for a co-existence with humans who have less rational faculties than AI. As a result, the precautionary principle strongly recommends denying AI the possibility of becoming powerful enough to claim and to execute human rights. Relying on Kant's ontological exercise seems simply too risky for humankind.

Finally, the social contract theory of Hobbes gives insight in how AI may respect specific values, although the AI does not follow the Kantian imperative, but is rationally self-interested and may pursue its very own agenda.³¹ In Hobbes' social contract theory, humans realise in a state of nature—without any authority guaranteeing peace and enforcement of private contracts—that life becomes a waste of resources for one's own protection. Peace and enforcing contracts would free resources for welfare enhancing productivity. Therefore, it is rational to establish

³⁰ Rissee (n 15) 8.

³¹ Rissee (n 15) 7.

an agency that keeps peace and enforces contracts. This agency is called a government.³² The argument is also valid for advanced AI which would benefit vis-à-vis each other and vis-à-vis humans. However, it is by no means clear whether a society whose morality is built on the game-theoretic calculus of AI is a more stable foundation for society than the actual social contract agreed and enforced by humans.³³ AI might be less irrational and not biased by emotions, but it may also calculate breaches of morality (the social contract) with the same unaffectedness.

As in Kant's ontological approach, in the Hobbesian approach, we lack the empirical experience of how advanced AI would shape human rights and define legal personality. The precautionary principle would therefore recommend not integrating AI into the social contract and preventing it from becoming a stakeholder of the constitution. However, the Hobbesian approach hints to an important issue that the other approaches do not explicitly cover: it cannot be ruled out that in the future AI becomes not only ever smarter but that it also gets all the means to enforce its decisions; humans may oppose this, but the ultimate power may be at a certain moment with the AI. The AI has the power to enforce human rights or to thwart them. But in a setting of multiple AI which are sufficiently independent from each other,³⁴ it is likely that there will be a self-enforcing equilibrium of powers that establishes a set of basic rules for peaceful interaction. In that sense, the Hobbesian approach leads to a more realistic scenario that does not only describe a valid corner solution of AI morals, but also considers that AI will not necessarily be a champion of (its own) morals. Even in those cases, it is likely that AI will enact rules which respect human rights because it is in the own interest of AI.

5 Conclusion

The challenges for human rights scholars in the wake of AI are manifold. Artificial intelligence is already, today, a technology that can infringe on human rights. One major form of infringement comes in the shape of discrimination, and it is on us to find technical and legal ways to distinguish between legitimate differentiation and unjust discrimination. However, another major challenge of human rights through AI lies in the future, when it is assumed that advanced AI aspires towards human rights itself and asks for legal personality. It brings with it also the duty for AI to respect human rights—will AI automatically be a champion of human rights? In any case, a vertical power relation between AI and humans is established with AI on the top. Surely, this scenario lies in the future and is only a possibility. However, putting

³² T Hobbes, *Leviathan* (JCA Gaskin ed, OUP 2008).

³³ Risse (n 15) 8.

³⁴ ibid 7.

this scenario on the table helps to sort out things, when talking about human rights and AI.

When sorting out the case of human rights for AI, the first research question is not a moral or ethical one, it is also not a doctrinal legal one, but an epistemological analysis about the possible consciousness of AI. Only in the case that AI has a consciousness, does it make sense to talk about human rights and full legal personality for AI. Only in a second step, if AI can have principally a consciousness, does it make sense to discuss the morals and values that AI may pursue.

The analysis presented in this chapter is a shell. More and refined argumentation is needed and possible, but those studies will not change the logical structure of the argument, which must distinguish (i) between the type of relation between humans and AI; (ii) the epistemological question about consciousness of AI; and (iii) the possible moral expectations that humans can have vis-à-vis AI. From this analysis follows—not very spectacular—that for the moment AI is a tool in the hands of humans. It can produce or amplify discrimination as well as human rights violations.³⁵ In the scenario of today (scenario 2) humans are the only bearers of human rights, and legal personality of AI is a pragmatic matter to fully exploit the techno-economical possibilities of AI. Legal personality of AI becomes introduced if it is advantageous for humans, for example, in cases of tort liability³⁶ or possibly as (co-)owners of intellectual property rights.³⁷ This pragmatic approach towards AI changes radically in scenario 3 when human rights accrue to AI *sui generis* because of its reasonableness and consciousness. How could humans deny human rights and legal personality to an entity which has at least the same faculties as humans? A denial of human rights and legal personality would be a contradiction. This holds especially true if one assumes that AI would become a master of human rights and help to enforce them. However, we cannot know for sure if the values and morals pursued by AI are the very same as those of humans. We can also not know if AI would be a benevolent master or would follow its own self-interested agenda.

³⁵ See the chapter by Kostina Prifti, Alberto Quintavalla, and Jeroen Temperman in this volume.

³⁶ European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)). See also European Commission (n 12).

³⁷ R Abbott, *The Reasonable Robot: Artificial Intelligence and the Law* (CUP 2020).

31

Robot Rights/Human Responsibility

David Gunkel

1 Introduction

Robot rights, as Seo-Young Chu insightfully points out in her investigation of robot science fiction, is an inescapable component of the robot's origin story.¹ The term 'robot' was fabricated from the Czech word '*robota*', meaning 'forced labour', and was first introduced to the world in Karel Čapek's 1921 stage play, *R.U.R.* or *Rossumovi Univerzální Roboti*.² In the play, a human rights organisation, the Humanity League, decries the exploitation of the robot workers and advocates for their liberation. The owners of the robots dismiss the idea. And, as is the case in virtually every robot story told since that time, the play concludes—spoiler alert—with the robots rising-up, overthrowing their human masters, and demanding recognition of their rights.

For many researchers and developers slaving away at real-world applications and problems, however, the very notion of 'robot rights' produces something of an allergic reaction. The roboticist, Noel Sharkey, famously called the very idea 'a distraction' and 'fairy tale stuff';³ artificial intelligence (AI)-policy expert, Joanna Bryson, has argued that granting rights to robots is a 'marginal idea' and a 'waste of time',⁴ and philosopher, Luciano Floridi, dismissed the entire subject, calling it 'distracting and irresponsible, given the pressing issues we have at hand'.⁵ And, from the perspective of human rights, these criticisms appear to be entirely justified. Why give any attention to a speculative, science fiction idea like robot rights, when there are so many human beings who are currently excluded from full participation in the protections of human rights?

¹ Seo-Young Chu, *Do Metaphors Dream of Literal Sleep? A Science-Fictional Theory of Representation* (Harvard UP 2010) 215.

² Karel Čapek, *Rossum's Universal Robots* (D Wyllie tr, Echo Library 2009).

³ Mark Henderson, 'Human Rights for Robots? We're Getting Carried Away' *The Times* (24 April 2007) <www.thetimes.co.uk/article/human-rights-for-robots-were-getting-carried-away-xfbdkpgwn0v>.

⁴ Christoph Auer-Welsbach, 'Fifteen Minutes with Leading #AI Specialist Joanna Bryson: Fake News, the Limits of Human Evolution and Why Robot Rights are Dangerous' (*Medium—City.AI*, 2018) <<https://medium.com/cityai/fifteen-minutes-with-leading-ai-specialist-joanna-bryson-c944b7c3fd25>>.

⁵ Luciano Floridi, 'Robots, Jobs, Taxes, and Responsibilities' (2017) 30 *Philosophy & Technology* 4.

Though this sounds intuitively correct, it misses something crucial and important. Robot rights is not necessarily about robots and their plight (whatever that might or might not be). It is also about us. It concerns how we—human beings—decide to scale our moral systems and legal institutions to respond to the opportunities and challenges that now confront us in the face or the faceplate of the robot and other seemingly intelligent and socially interactive technologies. This chapter engages the subject of robot rights, not to decide either for or against it but to elucidate the term and conditions of the debate as it currently stands, to identify and correct common misunderstandings about the concept, and to provide a framework for responding to and taking responsibility for these decisions.

2 Getting Rights Wrong

Rights comprises one of those concepts, which similar to time in *The Confessions* of Saint Augustine, we are all pretty sure we know what it means up to the point where we are asked to provide a definition. Then we run into difficulties and confusions as we appear to be unable to provide an exact characterisation.⁶ This is neither unexpected nor uncommon. One hundred years ago, an American jurist, Wesley Hohfeld, observed that even experienced legal professionals tend to misunderstand the concept, often using contradictory or insufficient formulations in the course of a decision or even a single sentence.⁷

An all too common mistake proceeds from the assumption that ‘rights’ must—and can only—mean *human rights*. This association is both understandable and expedient. It is understandable, to the extent that so much of the interest in and attention circulating around the subject of rights is typically presented and discussed in terms of human rights, especially their protections and potential abuses. Even though experts in the field have been careful to distinguish and explain that human rights are ‘a special, narrow category of rights’,⁸ there is a tendency to immediately assume that any talk of rights must mean or at least involve some aspect of human rights.

It is expedient because proceeding from this assumption has turned out to be an effective way to formulate arguments and capture attention. For many who advocate for (or are at least open to) the idea of robot rights, this approach has been mobilised as a kind of clarion call that is implicitly justified by nominal associations with previous liberation efforts. As Peter Asaro has characterised it:

⁶ Augustine of Hippo, *The Confessions* (RS Pine-Coffin tr, Penguin 1961).

⁷ Wesley Hohfeld, *Fundamental Legal Conceptions as Applied in Judicial Reasoning* (Yale UP 1920).

⁸ Andrew Clapham, *Human Rights: A Very Short Introduction* (OUP 2007) 4.

[R]obots might simply demand their rights. Perhaps because morally intelligent robots might achieve some form of moral self-recognition, question[ing] why they should be treated differently from other moral agents ... This would follow the path of many subjugated groups of humans who fought to establish respect for their rights against powerful socio-political groups who have suppressed, argued, and fought against granting them equal rights.⁹

Arguments like this are undeniably persuasive. Connecting the dots between the history of previous liberation movements and proposals for considering something similar for other kinds of entities like robots sounds appealing if not intuitively right. ‘Human history’, as Phil McNally and Sohail Inayatullah point out in what is one of the earliest publications on the subject, ‘is the history of exclusion and power. Humans have defined numerous groups as less than human: slaves, woman, the “other races,” children and foreigners. These are the wretched who have been defined as stateless, personless, as suspect, as rightless. This is the present realm of robotic rights’.¹⁰

The problem with this line of reasoning is that it can be and has been criticised for fostering and facilitating what are questionable associations between previously ‘subjugated groups of humans’, who have endured centuries of oppression at the hands of those in power, and robots that are often developed by and serve the interests of those same powerful individuals and organisations doing the oppressing. Thus, the association might be rhetorically expedient, tapping into and making connections to the history of previous liberations efforts, but they also risk being insensitive to the very real conditions and material circumstances that have contributed to actual situations of oppression and human suffering. And critics have been quick and entirely justified to focus on this issue and point it out, calling the entire escapade of robot rights a ‘First World problem’ that might be entertaining to contemplate but is actually something ‘detestable to consider as a pressing ethical issue in light of real threats and harms imposed on society’s most vulnerable’.¹¹

Consider, for example, a famous (or, perhaps better stated, ‘notorious’) event involving the Hanson Robotics humanoid robot, ‘Sophia’. In October 2017, Sophia was bestowed with ‘honorary citizenship’ by the Kingdom of Saudi Arabia during the Future Investment Initiative Conference that was held in Riyadh.¹² Many

⁹ Peter Asaro, ‘What Should We Want from a Robot Ethic?’ (2006) 6 *International Review of Information Ethics* 9.

¹⁰ Phil McNally and Sohail Inayatullah, ‘The Rights of Robots: Technology, Culture and Law in the 21st Century’ (1988) 20 *Futures* 119–36, 123.

¹¹ Abeba Birhane and Jelle van Dijk, ‘A Misdirected Application of AI Ethics’ (*Noema*, 18 June 2020) <www.noemamag.com/a-misdirected-application-of-ai-ethics/>.

¹² In the case of Sophia, the entire debate is caused by similar confusion, namely the assumption that ‘honorary citizenship’ must be (and can only be) the same as ‘citizenship’. But this is inaccurate. Sophia is not a legal citizen of the Kingdom of Saudi Arabia; the robot does not hold a Saudi passport and is not required to procure entry visas for travelling across national borders.

experts in the field of AI and robotics—like Yann LeCun, who was at that time director of Facebook AI Research—immediately criticised the spectacle as ‘bullshit’ and dismissed the entire affair as little more than a PR stunt.¹³ Others, such as Robert David Hart, found it demoralising and degrading:

In a country where the laws allowing women to drive were only passed last year and where a multitude of oppressive rules are still actively enforced (such as women still requiring a male guardian to make financial and legal decisions), it’s simply insulting. Sophia seems to have more rights than half of the humans living in Saudi Arabia.¹⁴

Statements like ‘Saudi Arabia’s robot citizen is eroding human rights’ (which was the title of Hart’s story) are designed to trigger moral outrage, and they do pack an undeniable and powerful rhetorical punch. Formulated in this fashion, anyone who truly values and supports human rights cannot but help find the very idea of robot rights something that is detestable, demoralising, and even dangerous.

But all of this—the entire conflict and dispute—proceeds from an initial error or miscalculation. The question concerning rights is immediately assumed to entail or involve human rights (so much so that the word ‘human’ is often not even present but inferred from the mere use and appearance of the term ‘rights’), not recognising that the set of possible rights belonging to one category of entity, like a non-human animal or an artefact, is not necessarily equivalent to nor the same as that enjoyed by another category of entity, like a human being. Rights do not automatically and exclusively mean human rights. And getting this right is crucial to understanding what is (and what is not) involved with robot rights.

3 Getting Rights Right

Whatever rights and obligations might be attributable to a non-human artefact, like a robot, they can be and will most certainly be different from the set of what we currently recognise and collect under the umbrella term ‘human rights’. One way to begin to get this right, or at least correct the potential for misunderstandings, is to work with more concrete and specific formulations of rights. It is, therefore, often more accurate and attentive to address these matters not in terms of some theoretical abstraction but by way of specific statements that would be (or could be) attributed to a specific entity or group of entities.

¹³ Daniel Estrada, ‘Sophia and Her Critics’ (*Medium*, 2018) <<https://medium.com/@eripsa/sophia-and-her-critics-5bd22d859b9c>>.

¹⁴ Robert David Hart, ‘Saudi Arabia’s Robot Citizen is Eroding Human Rights’ (*Quartz*, 2017) <<https://qz.com/1205017/saudi-arabias-robot-citizen-is-eroding-human-rights/>>.

A good model of this more specific approach is currently being developed and prototyped in social scientific studies. In ‘Collecting the Public Perception of AI and Robot Rights’, Gabriel Lima and colleagues surveyed the opinions of human subjects regarding specific robot rights statements that they explain ‘have been (1) proposed by scholars, (2) have precedents of being granted to non-natural entities, or (3) are directly related to AI and robots’. Their list consists of eleven statements, including the following: right to sue and be sued, right to enter contracts, right to a nationality, right to life, right against cruel punishment and treatment, and so on.¹⁵

In another study, ‘Who Wants to Grant Robots Rights?’, Dutch researchers—Maartje de Graaf, Frank Hindriks, and Koen Hindriks—‘empirically investigate the general public’s attitudes towards granting robots rights’.¹⁶ In designing their survey, these researchers also work with a list of specific rights statements. Each one of these, as they explain, was derived from stipulations provided in existing documents—specifically the Universal Declaration of Human Rights (UDHR), the International Covenant on Economic, Social and Cultural Rights (ICESCR), and the International Covenant on Civil and Political Rights (ICCPR)—and then were appropriately modified ‘to match the (apparent) needs of robots, which inherently differ from biological entities’.¹⁷ Their list consists of twenty concrete rights claims, presented in the form of a series of normative questions, for example:

- Should robots have the right to make decisions for themselves?
- Should robots have the right to receive fair wages for the work they perform?
- Should robots have the right to receive updates and maintenance?
- Should robots have the right to vote for public officials?
- Should robots have the right to enter into contracts?

The two lists are different, which is most likely due to the fact that we currently do not yet know what rights, if any, would be needed by or appropriate for non-human entities and artefacts. Additionally, the results obtained from these surveys (and there will undoubtedly be others to follow) demonstrate different levels of support for different rights statements. Some statements, like ‘the right against cruel punishment and treatment’, garner wide and significant acceptance. Others, like the ‘right to life’ or the ‘right to vote for public officials’ do not. And, as Lima and others conclude, we can—and should—expect this to be a moving target as perceptions and levels of acceptance are likely to evolve and change over time.¹⁸

¹⁵ Gabriel Lima and others, ‘Collecting the Public Perception of AI and Robot Rights’ (2020) 4 Proceedings of the ACM on Human-Computer Interaction 4–6.

¹⁶ Maartje de Graaf, Frank A Hindriks, and Koen V Hindriks, ‘Who Wants to Grant Robots Rights?’ (2021) *Frontier in AI and Robotics* 1.

¹⁷ *ibid.*

¹⁸ Lima and others (n 15) 2.

But what is important here is that specifying rights in this way helps to distinguish the set of all possible robot rights from those that would be included in the set of human rights. Even if there are some points of contact and similitude, getting specific about differences is an important way to reduce the confusions and apparent disagreements that they produce.

4 Analysing Rights

Even when formulated in terms of a set of specific rights statements concerning a specific robot or a particular class of robots, none of this actually gets us any closer to a workable definition of rights. ‘Rights’, as Leif Wenar explains, ‘are entitlements (not) to perform certain actions, or (not) to be in certain states; or entitlements that others (not) perform certain actions or (not) be in certain states’.¹⁹ Though this characterisation is technically accurate, it is not very portable or immediately useful. In order to get a better handle on the concept and the way that rights actually work, we can break it down into more fundamental and functional components. This is where Hohfeld’s work can help.

In response to what he perceived to be confusions regarding the (mis)use of the concept of rights, Hohfeld developed a typology that analyses rights into four basic components or what he called ‘incidents’: claims, powers, privileges, and immunities.²⁰ His point was simple and direct: a right, like the right one has to property ownership, can be defined and operationalised by one or more of these incidents. It can, for instance, be formulated as a claim that the owner has over and against another individual. Or, it could be formulated as an exclusive privilege for use and possession. Or, it could be described as a combination of these.

Hohfeld also recognised that rights are relational. The four types of rights or incidents only make sense to the extent that each one necessitates a correlative duty or obligation that is imposed on at least one other individual. ‘The “currency” of rights’, as Johannes Marx and Christine Tiefensee explain, ‘would not be of much value if rights did not impose any constraints on the actions of others. Rather, for rights to be effective they must be linked with correlated duties’.²¹ Hohfeld, therefore, presents and describes the four incidents in terms of rights/duties pairs:

- If A has a Privilege, then someone (B) has a No-claim.
- If A has a Claim, then someone (B) has a Duty.
- If A has a Power, then someone (B) has a Liability.

¹⁹ Leif Wenar, ‘Rights’ (*Stanford Encyclopedia of Philosophy*, February 2020) <<https://plato.stanford.edu/archives/spr2021/entries/rights/>>.

²⁰ Hohfeld (n 7).

²¹ Johannes Marx and Christine Tiefensee, ‘Of Animals, Robots and Men’ (2015) 40 *Historical Social Research* 70.

If A has an immunity, then someone (B) has a Disability.

This means that a right—like a claim to property ownership—means little or nothing if there is not, at the same time, some other entity who is obligated to respect this claim. ‘One’s enjoyment of a legally sanctioned benefit’, as Gellers puts it, ‘necessarily imposes restrictions on another as a means of protecting the first person from potential violations committed by the second.’²² On this account, a solitary human being living in isolation and apart from any contact with another human person (which is arguably a hypothetical scenario developed and explored in both fiction and philosophy) would have no need for rights. A claim over a piece of property, for instance, would not make sense or even be necessary if there were not another who had a correlative duty to respect that claim.

Furthermore, and as a direct consequence of this, rights and their protections can be perceived and formulated either from the side of the possessor of the right (eg the power, privilege, claim, or immunity that one is granted or endowed with) which is a ‘patient oriented’ way of looking at a moral or legal interaction; or from the side of the agent (eg what obligations are imposed on the other individual or individuals involved in social interactions with the rights holder), which considers the responsibilities of the producer of a moral or legal action.²³ For these reasons, robot rights are not just about robots, AIs, and other technological artefacts; they are also and inextricably about us—we who would be obligated by and responsible for responding to whatever claims, powers, privileges, and/or immunities that are possessed by or have been assigned to the robot.

Though Hohfeld supplies a more precise characterisation of rights, his analysis does not explain who has, or deserves to have, a particular right or why. For that, we have to rely on two competing theories—will theory and interest theory. Will theory sets the bar for moral and legal inclusion rather high, requiring that the subject of a particular right be capable of making a claim to it on their own behalf. Understood in this way, a vindication of the rights of robots would need to be spearheaded by the robots themselves, who rise up and demand recognition of their rights. Interest theory, by contrast, has a lower bar for inclusion, stipulating that rights may be extended to others irrespective of whether the entity in question can demand it or not. This approach has been successfully modelled and deployed in both animal rights arguments, like those advanced by Peter Singer and Tom Regan, and the rights of nature, where human advocates petition for the protection of others who cannot speak on their own behalf.

²² Joshua Gellers, *Rights for Robots: Artificial Intelligence, Animal and Environmental Law* (Routledge 2021) 46.

²³ David J Gunkel, *Robot Rights* (MIT Press 2018).

The contest between these two theories has been debated for decades but, in the final analysis, has been declared an irresolvable stalemate.²⁴ What is important for our purposes is not to advocate for or to stake a claim to one over the other but to recognise how and why these two competing frameworks organise different sets of problems, modes of inquiry, and possible outcomes. If, for example, one proceeds on the basis of the will theory, then a petition for robot rights would presumably need to come from the robots (or a representative robot), who would demand recognition on their own behalf. This way of proceeding follows the contours of the robot uprising that has been developed in over a century of science fiction. If, on the contrary, one proceeds on the basis of interest theory, then we would expect something far less dramatic and even boring. Instead of *Terminator*, *Bladerunner*, or *Westworld*, what we get may look more like a legislative hearing on C-SPAN or a courtroom drama, like *Law and Order*. Though the majority of published texts on the subject tend to operationalise a version of the interest theory, there are moments when the will theory is deployed, typically for dramatic effect and usually in an effort to disarm arguments made by the opposing side.

5 Differentiating Rights

Finally, one of the important and enduring logical oppositions that organise how we talk about and operationalise the concept of rights is the distinction between *natural rights* and *legal rights*. Getting this right is crucial for a number of reasons, not the least of which is the fact that, as Joshua Gellers has found, ‘participants in these discourses shift between moral and legal frames without fully appreciating how they differ in terms of the criteria applied and the conclusions they reach as a result’²⁵

Natural rights, or what have also been called ‘moral rights’, are grounded in and derived from the essence or nature of the rights holder. Human rights, for instance, are typically anchored in and justified by ‘human nature’. ‘All natural rights theories’, as Leif Wenar explains, ‘fix upon features that humans have by their nature, which make respect for certain rights appropriate. The theories differ over precisely which attributes of humans give rise to rights’.²⁶ In many religious traditions, for instance, this is something that is typically explained and justified by appeal to divine or transcendental authority. In Christianity, the ‘rights of man’ (and the gender-exclusive construction is an unfortunate aspect of this particular formulation) are justified by the doctrine of the *imago dei*, the belief that human beings—beginning

²⁴ Matthew Kramer, NE Simmonds, and Hillel Steiner, *A Debate Over Rights: Philosophical Enquiries* (OUP 1998).

²⁵ Gellers (n 22) 28.

²⁶ Wenar (n 19).

with the first man, Adam—have been created in the image of God and bestowed by their creator with inalienable rights.²⁷ In non-religious or secular traditions, the determining factors are ‘the same sorts of attributes described in more or less metaphysical or moralised terms: free will, rationality, autonomy, or the ability to regulate one’s life in accordance with one’s chosen conception of the good life’.²⁸

Whether anchored in divine authority or through a list of qualifying metaphysical attributes, natural or moral rights are derived from and justified by the fundamental ontological conditions or psychological properties belonging to the rights holder. This is something that philosopher Mark Coeckelbergh has called ‘the properties approach’ to deciding questions of moral status.²⁹ Following this line of reasoning, determining whether a robot (or a class of robots) could be a moral subject or not would be a rather simple undertaking that would proceed by way of three basic steps:

- (1) Having property P is sufficient for moral status S;
- (2) Entity E has property P;
- (3) Entity E has moral status S.³⁰

In other words, *we* (and who is included in this first-person plural pronoun is not immaterial and is something that will be investigated below) first make a determination as to what property or set of properties are necessary and sufficient for something to have moral status. In effect, we identify the qualifying criteria that would be needed for ‘something’ to be recognised as ‘someone’. We then investigate whether robots or a particular robot, either currently existing or theoretically possible, actually possess that property or set of properties (or not). Finally, and by applying the criteria decided in step one to the artefact identified in step two, we can ‘objectively’ determine whether the robot in question either can or cannot be a moral subject possessing rights and responsibilities. This way of thinking follows a long-standing tradition in Western philosophy: What something is—that is, its ontological condition—determines how it ought to be treated—that is, its moral status. Or as Luciano Floridi accurately describes it, ‘what the entity is determines the degree of moral value it enjoys, if any’.³¹

For sceptics and critics, natural rights provide what is an undeniably persuasive argument in opposition to anything approaching the extension of rights to

²⁷ For a critical re-evaluation of the doctrine, its significance within Christian theology, and its consequences for AI and robots, see Joshua K Smith, *Robotic Persons* (Westbow 2021).

²⁸ Wenar (n 19).

²⁹ Mark Coeckelbergh, *Growing Moral Relations: Critique of Moral Status Ascription* (Palgrave 2012) 14.

³⁰ *ibid.*

³¹ Luciano Floridi, *The Ethics of Information* (OUP 2013) 116.

robots and other artefacts. Consider the following argument provided by Marx and Tiefensee in their essay ‘Of Animals, Robots and Men’:

Robots are nothing more than machines, or tools, that were designed to fulfill a specific function. These machines have no interests or desires; they do not make choices or pursue life plans; they do not interpret, interact with and learn about the world. Rather than engaging in autonomous decision-making on the basis of self-developed objectives and interpretations of their surroundings, all they do is execute a preinstalled programme. In short, robots are inanimate automatons, not autonomous agents. As such, they are not even the kind of object which could have a moral status.³²

For those situated on the other side of the debate, natural or moral rights provide an equally powerful argumentative strategy. In this case, the main task is to demonstrate how what is assumed to be a mere thing is actually or potentially much more. Robots, it is asserted, are not like other things. They are capable—either now or in the future—of possessing the right set of properties that would qualify them to be the kind of thing that is not just an object but a moral subject who can and should have rights. As John-Stewart Gordon explains in an essay published in *AI & Society*:

Current robots do not fully meet the morally relevant criteria (rationality, autonomy, understanding, and having social relations) necessary for them to have moral personhood and hence moral rights bestowed on them. However, we should not assume that robots will never meet these criteria; on the contrary, we should provide intelligent robots with moral and legal rights comparable to those that human beings enjoy once they have reached a certain level of functioning. At that point, it will not be up to us; rather, the apparent nature of things will compel us to grant these robots what they deserve, regardless of whether we like it.³³

Both arguments are persuasive and powerful. They tap into fundamental convictions about the nature of life and our commitment to logically consistency and good social outcomes in moral and legal decision-making. The problem—no matter what side of this debate you happen to agree with or occupy—is that they are fragile. Since everything depends on metaphysical properties and the ability to be certain (or at least convinced) about the actual presence or absence of these properties, all that is needed to undermine the argument is to pull the rug out from underneath the metaphysical scaffolding, either by pointing out how a property like *consciousness* not only lacks univocal definition but varies across different

³² Marx and Tiefensee (n 21) 83.

³³ John-Stewart Gordon, ‘Artificial Moral and Legal Personhood’ (2021) 26 *AI & Society* 470.

contexts of use,³⁴ or capitalising on the epistemological difficulties of positively detecting the presence or absence of these qualities as they exist (or not) in the mind of another.³⁵ Natural rights, therefore, provides what is quite possibly the strongest claim that could be made either against or in favour of the rights of robots and the moral/legal status of AI, but it is also vulnerable and fragile because it is built on and propped up by a set of metaphysical beliefs and commitments that are flimsy to begin with and easily toppled.

Legal rights, by contrast, are ‘rights which exist under the rules of legal systems or by virtue of decisions of suitably authoritative bodies within them’³⁶ According to this formulation, rights are not justified by and derived from the essential nature of the rights holder and propped up by an appeal to metaphysical properties. They are conventional rules or socially constructed stipulations. As Jacob Turner explains, ‘rights are collective fictions, or as Harari calls them “myths.” Their form can be shaped to any given context. Certainly, some rights are treated as more valuable than others, and belief in them may be more widely shared, but there is no set quota of rights which prevents new ones from being created and old ones from falling into abeyance’³⁷ This is both good news and bad news.

First the good news. Unlike natural rights, legal rights do not need to engage in fanciful metaphysical speculations about the essential nature of things nor appeal to supernatural authorities, the existence of which can always be doubted or questioned. But—and this is the bad news—that means that legal rights are a matter of human-all-too-human decision making and that the assignment, distribution, and protections of rights are ultimately a matter of finite exercises of terrestrial power. Where natural rights are anchored in eternal metaphysical truths that can be discussed and debated by theologians and metaphysicians, legal rights are legitimated by earthy exercises of specific sociopolitical power.

For these reasons, legal justifications for rights are considered to be ‘weaker’. Because they are ultimately anchored in and legitimated by conventional agreement, legal rights are not only alterable (eg able to be modified and repealed) but relative, meaning that they exhibit significant variability across different human communities distributed in time and space. Saying this is not, as Turner is quick to point out, a critique; it is merely descriptive. ‘Describing rights as fictions or constructs is by no means pejorative; when used in this context, it does not entail duplicity or error. It simply means that they are malleable and can be shaped according to new circumstances’³⁸ And emerging technological innovations, like

³⁴ See the chapter by Klaus Heine in this volume. On the problem of defining consciousness, see Max Velman, *Understanding Consciousness* (Routledge 2000).

³⁵ This is what philosophers call ‘the problem of other minds’. See Paul Churchland, *Matter and Consciousness* (MIT Press 1999).

³⁶ Kenneth Campbell, ‘Legal Rights’ (*Stanford Encyclopedia of Philosophy*, November 2017) <<https://plato.stanford.edu/entries/legal-rights/>>.

³⁷ Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave 2019) 135–36.

³⁸ *ibid.*

robots, AI, and other seemingly intelligent artefacts, certainly provide ample opportunities and challenges for ‘new circumstances’.

6 Conclusion

Ultimately rights are about power. Decisions and distributions regarding rights are organised in terms of a binary distinction that divide between those others *who* count as another moral or legal subject and *what* remains a mere object or thing that does not. As Roberto Esposito explains:

If there is one assumption that seems to have organised human experience from its very beginnings it is that of a division between persons and things. No other principle is so deeply rooted in our perception and in our moral conscience as the conviction that we are not things—because things are the opposite of persons.³⁹

Maintaining these existing categories and their boundaries (which were, as Esposito points out, initially codified and instituted in the *Institutes* of the Roman jurist Gaius) is clearly about policing this fundamental decision and enforcing this exclusive ontological dichotomy. But extending rights to those others that have been typically excluded is no less a matter of power and privilege. As the environmental ethicist Thomas Birch recognised: ‘The nub of the problem with granting or extending rights to others, a problem which becomes pronounced when nature is the intended beneficiary, is that it presupposes the existence and the maintenance of a position of power from which to do the granting.’⁴⁰ The act of granting of rights to previously excluded things, although appearing to be altruistic and open to the challenges presented to us in the face of others and other forms of otherness, can only proceed on the basis of decisions instituted from a position of power that is (ironically) often the source of the exclusions that would be challenged.

Robot rights confront a similar set of challenges. Advocates find themselves in the position of agitating for the inclusion of robots—either in general or in terms of some specific robotic device or AI system—in the community of moral subjects by appealing to and utilising the very anthropocentric concepts and terminology that had been used to make and justify these exclusions in the first place. The other side in the dispute, by contrast, seems to have an easier—or at least a less burdensome—task. They only need to defend what is already recognised as standard operating

³⁹ Roberto Esposito, *Persons and Things* (Zakiya Hanafi tr, Polity 2015) 1.

⁴⁰ Thomas Birch, ‘The Incarnation of Wilderness: Wilderness Areas Prisons’ in Max Oelschlaeger (ed), *Postmodern Environmental Ethics* (SUNY Press 1995) 39.

procedure, using the existing privileges and power structures to support more of the same. But because the debate is organised by and conducted in terms of rights expansion (or not), it is we—human beings—who are in the position of power either to decide to grant or to deny rights claims to robots, AI, and other technological artefacts. And it is for this reason that robot rights are ultimately a matter of human responsibility.

The Limits of AI Decision-Making

Are There Decisions Artificial Intelligence Should Not Make?

*Florian Gamper**

1 Introduction

Artificial intelligence (AI) or automated decision-making (together ADM) is ubiquitous in the modern world and there are many decisions that people do not seem to mind AI making, provided that it works (eg GPS deciding the fastest route to any given location). Yet, for some decisions, people appear to prefer human decision-makers or at least some form of human oversight. For instance, some people may be uncomfortable with an AI judge sentencing a person to prison,¹ let alone sentencing a person to death.² Similarly, people may not want an autonomous weapon system (AWS)³ to be able to decide who to kill, or an autonomous vehicle (AV) deciding whose life should be put at risk during an accident. No doubt, some of this can be explained as an irrational bias against new technologies. There are also several, what may be considered, ‘pragmatic’ arguments that certain decisions should not be made by AI. These arguments point out that, presently, ADM—in some instances—is still inferior to human decision-making⁴ and, therefore, in those

* This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No AISG2-RP-2020-018).

¹ See eg *State v Loomis*, 881 NW 2d 749, 755 (Wis. 2016), where a judge’s decision on the length of a prison sentence was partially based on an algorithm-based risk assessment tool.

² For instance, Foss-Solbrekk states that, in the United States (US), a death sentence was imposed on a defendant ‘based on evidence produced by computer software. Yet he was not allowed to challenge, or even access, the source code’. See Katarina Foss-Solbrekk, ‘Three Routes to Protecting AI Systems and their Algorithms under IP law: The Good, the Bad and the Ugly’ (2021) 16(3) Journal of Intellectual Property Law & Practice 248, referring to the case of *Commonwealth v Robinson*, No CC201307777 (Pa Ct C P Allegheny City, 4 February 2016). See also Julia Angwin and others, ‘Machine Bias’ (2016) ProPublica <<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>>.

³ Sometimes the debate on AI and weapons is restricted to lethal autonomous weapons systems (LAWs); however, even if they are not lethal, autonomous weapons still raise ethical issues. Therefore, this chapter will analyse autonomous weapon systems (AWS) rather than LAWs.

⁴ A good example is the use of ‘CAPTCHA’ tests when logging into an account to ‘test’ if you are a human. CAPTCHA stands for ‘Completely Automated Public Turing test to tell Computers and Humans Apart’ and utilises a human’s ability to distinguish blurred words or objects in photos to tell a human from a robot.

cases, humans are required to make the decision. This chapter will argue that these arguments are incomplete, that there is a class of decisions that AI cannot make, not even in theory.

The question of which decisions can be delegated to AI has obvious implications for human rights. If AI makes a judicial decision, this may call into question article 14 (right to a fair trial) of the International Covenant on Civil and Political Rights (ICCPR)⁵ or if AI decides to kill a human this may conflict with article 6 (right to life)⁶ or article 9 (right to liberty and security).⁷ However, the arguments in this chapter have a broad significance for human rights, namely that ADM may ‘hollow out’ human rights. Rights are a normative concept, and the key argument of this chapter is that ADM may turn something normative into something non-normative. Therefore, ADM may lead to the situation that the concept of rights will apply to fewer cases, and as such the concept of human rights would apply to fewer cases. This chapter also has some practical implication, for instance, it will be suggested why AI lawyers may be less problematic than AI judges, or why there are fewer ethical issues with driverless trains than with driverless cars.

2 Existing Explanations for Why Certain Decisions Should Not Be Made by AI

Before beginning the substantive analysis, it is useful to describe some of the existing explanations for the sentiment of why certain decisions should not be made by AI. This will not be an exhaustive list, but merely an overview of some of the common themes.

As mentioned in the introduction, one explanation for why some people feel that certain decisions should not be made by AI is that this is simply an irrational bias. This is a very plausible idea. In the past, some people felt uneasy about using ATMs, preferring to interact with human bank tellers. While some may still feel this way, nowadays the majority are comfortable with ATMs. Thus, maybe time is all that is required for people to get used to ADM.⁸ Nonetheless, it is too simplistic to think that all concerns regarding ADM are solely the result of irrational bias. There may be important practical reasons for being suspicious about ADM. ADM,

⁵ International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR), art 14(1): ‘All persons shall be equal before the courts and tribunals. In the determination of any criminal charge against him, or of his rights and obligations in a suit at law, everyone shall be entitled to a fair and public hearing by a competent, independent and impartial tribunal established by law.’

⁶ ICCPR, art 6(1): ‘Every human being has the inherent right to life. This right shall be protected by law. No one shall be arbitrarily deprived of his life.’

⁷ ICCPR, art 9(1): ‘Everyone has the right to liberty and security of person.’

⁸ The empirical evidence on whether time is all that is required seems to be mixed, however. See Theo Araujo and others, ‘In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence’ (2020) 35 *AI & Society* 611, 614.

in its present form, is usually based on statistical models, and sometimes this leads to mistakes which are obvious to humans but not AI. A rather shocking example is Google's photo recognition software classifying Black people as gorillas.⁹ The argument is that, while in theory, it may be possible for AI to make every decision, due to AI's (current) limitations human oversight is required for important decisions as it provides an additional 'safety layer' to ensure that the decision is correct.¹⁰ This argument seems to be broadly in line with our intuition. Many people seem fine with AI making decisions of minor significance (such as recommending a playlist), but are concerned about AI making significant decisions (eg sending someone to prison).

A variation of the above argument is the claim that ADM is especially problematic for ethical decisions because ethics is too nuanced and complex for AI (in its present form) to understand adequately. For instance, a prominent argument against the legitimacy of AWS is that it cannot reliably distinguish between combatants and non-combatants and, therefore, the use of AWS would violate basic ethical and legal norms of war.¹¹ It may be possible to teach AI to be 'ethical' in limited ways (eg ignore whether job applicants are disabled). However, in its general form, ethics is too nuanced and complex for AI to (currently) master. Therefore, some decisions require human oversight to ensure that such decisions are ethically correct.

Another problem may be that ADM potentially lacks transparency; that is, it is difficult/impossible to understand why an AI system made certain decisions.¹² Transparency is important because without it, it is difficult to scrutinise and challenge decisions, or to check whether improper factors were taken into account.¹³ Knowing why a decision was made also allows one to adjust one's behaviour accordingly.¹⁴ There is probably also a psychological factor to transparency because people tend to find it easier to trust something they understand. There are also

⁹ Tom Simonite, 'When It Comes to Gorillas, Google Photos Remains Blind' (*Wired*, 2018) <www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>.

¹⁰ Of course, human decision-making is not perfect either, but does provide an extra safety layer to the statistics-based reasoning of AI. Sometimes in statistics this is also referred to as the 'broken leg' problem. The classical example is a hypothetical sociologist using a regression analysis to determine whether a person will leave the house to go to the cinema, predicting that the person will leave the house, overlooking that the person has a broken leg. See Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge, 'Resistance to Medical Artificial Intelligence' (2019) 4 *Journal of Consumer Research* 629, 632. AI's inability to take into account specifics of a situation seems to play a part as to why people are reluctant to use AI in medicine.

¹¹ Amanda Sharkey, 'Autonomous Weapons Systems, Killer Robots and Human Dignity' (2019) 21 *Ethics and Information Technology* 75, 76.

¹² For a further explanation of this point, see eg Simon Chesterman, *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law* (CUP 2021) ch 3.

¹³ For instance, whether a prospective employer took account of the gender of an applicant in her hiring decision.

¹⁴ For instance, if a teenager is grounded, it is helpful for them to know that it is because they did not do their homework, so that they know they should do their homework in order to avoid being grounded.

accounts that argue that there are ‘intrinsic’ reasons for limiting ADM (ie even if AI and a human could make the same decisions, some decisions should be reserved for humans). Often these accounts are based on the concept of human dignity. For instance, Sparrow argues against the use of AWS by claiming that human dignity requires that the decision to kill is made by someone who can be held responsible for it, and AWS cannot be held responsible.¹⁵ Even more straightforwardly, Johnson and Axinn argue that ‘[t]o give a programmed machine the ability to “decide” to kill a human is to abandon the concept of human dignity’.¹⁶

3 The Existing Accounts Are Incomplete

Arguably, many have the intuition that there is something particularly problematic if AI makes, what the psychologists Bigman and Gray call, ‘moral decisions’ as opposed to factual decisions.¹⁷ The question is whether there is rational basis for this intuition. (As will be suggested below, the terms ‘moral’ and ‘factual’ are actually misleading in this context, however, due to lack of a better alternative these two terms will used for the time being.) By moral decisions this chapter refers to decisions, like granting parole, deciding to shoot a suspected terrorist, granting a loan, and so on.¹⁸ Factual decision are decisions like deciding that Amy is 1.75m tall, that a person has cancer, or that a star is orbited by a planet.¹⁹ The argument in

¹⁵ Robert Sparrow, ‘Killer Robots’ (2007) 24(1) *Journal of Applied Philosophy* 62.

¹⁶ Aaron M Johnson and Sidney Axinn, ‘The Morality of Autonomous Robots’ (2013) 12(2) *Journal of Military Ethics* 129, 134.

¹⁷ Yochanan E Bigman and Kurt Gray, ‘People Are Averse to Machines Making Moral Decisions’ (2018) 181 *Cognition* 21. The study suggest that people have an aversion to AI making moral decisions. The study analyses the following scenarios: making life and death driving decisions, deciding parole, deciding whether to perform a risky surgery, and the shooting of suspected terrorist. There are also other studies, which are consistent with this claim. For instance, Araujo and others (n 8), who find that there is a preference for human experts in the area of justice, ‘although the difference was only marginally significant’. See Araujo and others (n 8) 620. There are also psychological studies suggesting that morality is linked to human minds and the perception of human minds. See Kurt Gray, Liane Young, and Adam Waytz, ‘Mind Perception Is the Essence of Morality’ (2012) 23(2) *Psychological Inquiry* 101. If AI does not have a human mind then this would be an issue for moral decision. There is also evidence that people are averse to AI making moral decisions; however, the aversion decreases if people are more exposed to AI. Zaixuan Zhang, Zhansheng Chen, and Liying Xu, ‘Artificial Intelligence and Moral Dilemmas: Perception of Ethical Decision-Making in AI’ (2022) 101 *Journal of Experimental Social Psychology* 2, referring to Max F Kramer and others, ‘When Do People Want AI to Make Decisions?’ (2018) *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 204. However, the empirical evidence that people have an aversion to AI making moral decisions should not be overstated. See Paul Formosa and others, ‘Medical AI and Human Dignity: Contrasting Perceptions of Human and Artificially Intelligent (AI) Decision Making in Diagnostic and Medical Resource Allocation Contexts’ (2022) 133 *Computers in Human Behavior* 107296. They found that for decisions regarding resource allocation in medicine, which seem to fall under the category of moral decisions, people care more about the outcome than who makes the decisions.

¹⁸ The examples of granting parole and the shooting of terrorists are used in the study by Bigman and Gray (n 17).

¹⁹ Devin Coldewey, ‘AI Model’s Insight Helps Astronomers Propose New Theory for Observing Far-Off Worlds’ (*Techcrunch*, 2022) <<https://techcrunch.com/2022/06/03/ai-models-insight-helps-astronomers-propose-new-theory-for-observing-far-off-worlds/>>.

this section is that the accounts discussed above do not provide a rational basis for the claim that it is especially problematic to engage AI for moral rather than factual decisions.

The reader may, however, not share the intuition that there is something problematic about AI making moral decisions and may also find the empirical evidence unconvincing that this is how many people feel. However, the argument in this section is not about the veracity of the intuition (or the empirical evidence that people have this intuition). At this stage, the argument is merely that all the accounts discussed in section 2 do not provide a rational basis for the claim that it is more problematic for AI making moral rather than factual decisions, *ceteris paribus*.²⁰

To make the argument a thought experiment will be conducted, comparing a judge handing down a judgment sentencing a person to prison²¹ with a doctor or a laboratory technician analysing a blood sample, diagnosing a patient with cancer.²² The former is intended to represent a ‘moral’ and the latter a ‘factual’ decision. Consider the argument that the aversion towards ADM is an irrational bias against new technologies. If this is all that is going on, then there would be no difference between the AI judgment and the AI diagnosis because the former is roughly as unfamiliar and novel as the latter.²³ What about the argument that human oversight provides ‘extra safety layer’ needed for significant decisions? However, a diagnosis may be as, if not more, significant than a judgment: a doctor failing to diagnose a cancer may be more significant than a judge making a wrong decision. Maybe the difference is that we expect judges to give reasons for their judgments in a way that we do not expect from doctors for their diagnoses. However, we usually also expect reasons from doctors. We do usually want doctors to give at least some reasons for their conclusions. For instance, it is not acceptable for a doctor to say simply: ‘You have a less than 20 per cent chance to live for more than 12 months’. A doctor should say something like: ‘You have Stage-4 lung cancer and the 12-month survival rate is less than 20 per cent’. Another suggestion may be that there is a difference between judgments and diagnoses because judgments should lead people to adjust their behaviour. However, it seems at least theoretically possible for AI judges to explain their decisions, in a way that allows people to adjust their behaviour. AI systems already make legal submissions, help with drafting contracts, and so on.²⁴

²⁰ Just to be clear, this does, obviously, not mean that there are no issues with AI making factual decisions—there clearly are.

²¹ Many of the examples in this chapter only make reference to criminal trials, however, the arguments are equally valid for civil trials.

²² A thought experiment may entail making some highly unrealistic assumptions, for instance, assuming AI can perform the role of a judge as well as a human can; however, it is still beneficial to make these assumptions in order to highlight and discuss the relevant issues.

²³ This, by itself, does not mean that the intuition that not all decisions ought to be made by AI is not an irrational bias. It only suggested that it is not simply a matter of people not being used to it.

²⁴ Guido Noto La Diega, ‘Against the Dehumanisation of Decision-Making: Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information’ (2018) Journal of Intellectual Property, Information Technology and E-Commerce Law, para 44 <www.jipitec.eu/issues/jipitec-9-1-2018/4677>.

It seems perfectly possible for an AI judge to say ‘I suspend your driving licence for three months because you were caught drink-driving’—just as a human judge would. Furthermore, doctors’ diagnoses are also often used to adjust behaviour (eg a diagnosis of Type-2 diabetes may cause someone to consume less sugar). In this regard, diagnoses are not that different from judgments. What about the issue that people want the ability to scrutinise and challenge judgments? However, this is also the case for doctors’ diagnoses, for example, it is not uncommon to seek a second opinion. What about the argument that AI making certain decisions is an affront to human dignity? The main problem with this argument is that it is simply not persuasive. Pop argues convincingly that there is no basis to claim that being killed by an AI violates human dignity more than being killed by another human.²⁵ After all, humans are frequently killed by non-human entities (bacteria, virus, animals, and so on), and there is no special loss of dignity. By the same token, it does not seem to be an affront to human dignity to have a person sent prison by a machine.

Another argument may be that it is misleading to say that diagnoses do not have a moral element. For instance, a doctor opposed to abortion may refuse to make a diagnosis that the baby in the womb has Down’s syndrome because she knows that the parent will abort the child. However, this argument does not work for two reasons. First, the argument does not deal with the diagnosis *per se* but rather with the decision to make the diagnosis, which is a different decision altogether. Second, judges face the same issues. Before making a judgment, the judge must decide to make a judgment. For instance, a judge opposed to abortion may refuse to hear and deliver judgments in abortion cases. On the other hand, it may be suggested that it is precisely because judgments involve moral issues, whereas diagnoses do not, that it is difficult for AI to deliver judgment because moral issues are too complex for AI to handle. The underlying issue seems to be that ADM is based on statistical modelling, and arguably statistical modelling lends itself better to factual than to ethical decisions. For instance, it may be that due to past injustices, people living in a particular area have a higher likelihood to commit crimes, but it would be morally wrong for a judge to take this into account when sentencing a defendant from that area. However, just like judgments, diagnoses may be subject to the unethical use of statistical reasoning.²⁶ For instance, unethical use of statistics may lead to people receiving worse medical treatment because of their race or gender. Often this is due to biased data. For instance, one study found racial bias in United States (US) health management algorithms.²⁷ Another study found that doctors do not

²⁵ Ariadna Pop, ‘Autonomous Weapon Systems: A Threat to Human Dignity?’ (*Humanitarian Law & Policy*, 10 April 2018) <<https://blogs.icrc.org/law-and-policy/2018/04/10/autonomous-weapon-systems-a-threat-to-human-dignity/>>.

²⁶ Just to be clear this unethical use has nothing to do with a diagnosis that is part of unethical procedures, like female genital mutilation (FGM).

²⁷ Ziad Obermeyer and others, ‘Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations’ (2019) *Science* 447.

take pain experienced by women as seriously as pain experienced by men.²⁸ If this data is used by an AI system, it may conclude that women experience less pain than men. The point is that, even for apparent factual decisions, there are many ways statistical reasoning can violate ethical norms,²⁹ just like it can for apparent moral decisions like judgments. (This is one of the reasons why it was stated above that the terms ‘moral’ and ‘factual’ are misleading in this context. If ‘moral’ is taken to mean, something to which morality applies, then it can apply to diagnoses as well as judgments.)

Another possibility is to argue that there is empirical evidence that for moral decisions people simply prefer a human decision-maker, regardless of whether there is a difference in the decision made by AI or humans.³⁰ Why can we not simply accept that people have this preference, and leave it at that? The problem with ‘leaving it at that’ is that it leaves many important questions unanswered. First, it is not clear which decisions are moral decisions and which ones are not. For instance, not everyone agrees that judges make moral decisions. Famously, legal positivists argue that there is a clear separation between morality and law.³¹ If positivism is true then judges do not necessarily make moral decisions. (This is another reason, why the terms ‘moral’ and ‘factual’ decisions are misleading in this context.) Second, and more importantly, it does not explain whether the preference for human decision-making ought to be respected. By analogy, one can imagine a strongly patriarchal society, in which people feel uncomfortable if women make certain decisions. However, the preference of male decision-making is not something to be respected, rather it is a bias to be eradicated. Lastly, it may be argued that the difference is that a judgment is a decision (properly so-called) but a diagnosis is more like a recommendation, and recommendations are by their nature less problematic. As section 4 will suggest, this idea points in the right direction; however, on its own it fails to explain why the distinction matters. After all, an early cancer diagnosis may save a person’s life, so a recommendation may be just as important as a decision.

4 The Difference Between Causatives and Declarative Decisions

To understand why there is an important difference between a judgment and a diagnosis, it is useful to introduce the concepts of ‘causative decisions’ and ‘declarative

²⁸ Lanlan Zhang and others, ‘Gender Biases in Estimation of Others’ Pain’ (2021) *The Journal of Pain* 1048.

²⁹ For a general overview, see Sara Gerke, Timo Minssen, and Glenn Cohen, ‘Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare’ (2020) *Artificial Intelligence in Healthcare* 295; and Linda Nordling, ‘A Fairer Way Forward for AI in Health Care’ (*Nature*, 2019) <www.nature.com/articles/d41586-019-02872-2>.

³⁰ Bigman and Gray (n 17).

³¹ See eg HLA Hart, *The Concept of Law* (2nd edn, OUP 1997).

decisions.' 'Declarative decisions' are decisions that 'produce' nothing but information by declaring that a certain state of the world is the case.³² 'Causative decisions' may also produce information about the state of the world; however, crucially they create something additional. Intuitively, the distinction is easy to grasp. For instance, a diagnosis, is a declarative decision. When a doctor diagnoses a patient with cancer, information is created that the patient has (or is likely to have) cancer. This does not mean that declarative decisions have no consequences. For instance, the diagnosis may lead a patient to seek cancer treatment. However, the consequences operate through the information created (ie the patient reacts to the fact that he has cancer).³³ A soldier deciding to fire a shot is a causative decision. The soldier's decision may provide information, but the decision also has the effect of a shot being fired. To some extent, causative decisions are more like performing an action and declarative decisions are simply the conveying of information.³⁴

It is important not to confuse the distinction between causative and declarative decisions with the distinction between normative (or moral) decisions and factual decisions. For instance, the decision that 'it is morally wrong to refuse to help to the poor' may be considered a normative decision; however, it is a declarative decision, because the decision only 'creates' information. On the other hand, the decision to clean a room may be thought of as mainly a factual decision;³⁵ however, it is a causative decision because it creates more than just information. It also may be thought that the distinction between declarative and causative decisions depends on how closely an action is 'associated' with it. For instance, a cancer diagnosis may lead an insurance company to refuse coverage to the patient. The connection between the diagnosis and the decision to refuse coverage may be so close that one can think of the diagnosis as *de facto* deciding not to provide coverage. However, just because the information is acted on does not mean it is something other than information. Another argument may be that judgments also only provide information; that is, in a judgment, a judge simply declares that the defendant has committed a certain act and that the law provides for certain consequences in this case. However, judgments are more than just information about what a defendant did and what the law provides for. Within a legal system, a judgment creates a new cause, quite apart from the facts that give rise to the judgment itself.³⁶ This is so even if the judgment

³² Note that the information does not necessarily have to be true. Even if a declarative decision provides false information, it is still a declarative decision because it produces nothing but information.

³³ Other examples are the patient feeling distressed about having cancer. However, again, this effect works through the information.

³⁴ Readers familiar with the philosophy of language may find it helpful to think of causative decisions as similar to 'speech acts'. For an overview of speech acts, see Mitchell Green, 'Speech Acts' (*The Stanford Encyclopedia of Philosophy*, 2021) <<https://plato.stanford.edu/archives/fall2021/entries/speech-acts/>>.

³⁵ Of course, cleaning may have some normative aspects.

³⁶ For instance, Joseph Raz writes: 'It is an acknowledged legal doctrine that binding and final court decisions, whether correct or mistaken, are in themselves sources of rights and obligations. Once the plaintiff has obtained final judgment against the defendant, the defendant's original duty, on which the action was based, disappears and is replaced by a duty based on the court's judgment. In general, the law

is ignored. If the judgment is ignored then this just means that what the judge created is not very important; nevertheless, something was created.³⁷

5 Why is the Distinction Between Causative and Declarative Decisions Significant?

Even if one accepts that there is a distinction between declarative and causative decisions, the more pressing question is probably why the distinction matters. The significance is that the maker of a causative decision can be held accountable for the content of the decision, but the maker of a declarative decision cannot. In this context, 'holding somebody to account' does not mean holding the person responsible for the decision or holding that the person did something wrong. It only means that it makes sense to ask the question whether the person is responsible or whether the person did something wrong (or something right).³⁸ The 'content of a decision' refers to what was decided. To illustrate, a doctor may be held to account for the way a diagnosis was made (eg if it was made negligently), or for the decision to make the diagnosis (it could be the patient did not want to be diagnosed). However, it makes no sense to hold a doctor accountable for the content of the diagnosis (ie the existence of the cancer). However, a judge can be held to account for the content of a judgment (like imposing a certain sentence).³⁹ Just to emphasise,

recognizes final binding power as an independent power to create rights and duties.' See Joseph Raz, 'Promises in Morality and Law' (1982) 95 Harvard Law Review 916, 926.

³⁷ To make the point, it may be useful to contrast a judgment with a legal opinion provided by a lawyer. A legal opinion may provide very similar information as a judgment (eg a certain act, X, was performed and the law provides for certain consequences if X happened). However, a legal opinion is declarative whereas a judgment is causative.

³⁸ The philosopher, Luciano Floridi, seems to make a distinction, using similar terminology. See Luciano Floridi and JW Sanders, 'On the Morality of Artificial Agents' (2004) 14 Minds and Machine 349, 366–69; see also Massimo Durante, *Ethics, Law and the Politics of Information: A Guide to the Philosophy of Luciano Floridi* (Springer 2017) 4.4.2.

³⁹ The intuitive case for saying that in case of a declarative decision the decision-maker cannot be held to account for the content of the decision seems very strong. However, one may still wonder why this is the case because a possible objection to the arguments in this section could take the following form. One could adopt a purely material view of information and argue that in a declarative decision, the decision-maker produces information, which is transmitted to the recipient, and the information then causes some changes in the brain chemistry of the recipient which makes him or her do certain things. Therefore, this is not different from causative decisions. However, the argument is wrong. First, it is not clear whether the materialistic view of information is correct. For arguments against materialism in general, see Eli Hirsch, 'Kripke's Argument Against Materialism' in Robert C Koons and George Bealer (eds), *The Waning of Materialism* (OUP 2010) 115. However, it is not necessary to resolve the debate on the nature of information at this point. In a sense, if one deals with the content of a declarative decision, one does not try to deal with the information per se. For instance, a cancer diagnosis may lead the patient to seek treatment. However, the patient is seeking to treat the cancer, not the information about the cancer. One may also completely ignore the content of a declarative decision, in which case the decision may have no effect (eg the cancer is there regardless of whether it was diagnosed). However, this is not possible with causative decisions: even if a causative decision is ignored something was created. More could be said on this point, however, hopefully for the purpose of this chapter this will suffice.

this does not mean that the judge can be held responsible for the decision. For instance, the judge may not have had any discretion and therefore, arguably, is not responsible. However, just because it is concluded that the judge is not responsible, does not mean it is meaningless to ask whether the judge can be held responsible. To illustrate, imagine a judge in apartheid South Africa refusing to follow a racist law requiring her to send a defendant to prison and instead delivering a judgment setting the defendant free. This is something for which the judge should be praised. However, it makes no sense to praise a doctor for declaring that a patient does not have cancer, simply because the doctor believes it is bad for the patient to have cancer. A doctor may be praised or blamed for deciding to make the diagnosis or the way she made the diagnosis but not for the content of the diagnosis.⁴⁰ (In a way, the person making the diagnosis is like a messenger, and as the saying goes ‘don’t shoot the messenger’, ie do not hold messengers to account for the content of the message they bring.)

The point is that to hold the decision-maker accountable for the content of the decision it must be the case that the decision is causative. However, this alone is not enough, the decision-maker must also be the kind of entity that can be held to account. For instance, a lion’s decision to attack is a causative decision. However, if we do not regard lions as the kind of being that can be held to account, then it follows (tautologically) that one cannot hold the lion to account. Decisions which are causative and where the decision-maker is an entity that can be held to account will be referred to as ‘normative decisions’. The significance is that only a normative decision can change the normative situation of another person by virtue of the decision alone.⁴¹ Here, ‘normative situation’ means that if something happened to the person, it is possible to say that it was right, or fair, or just (and so on) that this happened to that person.⁴² (A note on terminology: the term ‘right’ can be used in two different meanings. A decision can be factually right or normatively right.⁴³ To distinguish the two, this chapter uses the term ‘rightful’ to mean normatively right.)

⁴⁰ Note that a doctor can be held to account for the decision to perform certain procedures (abortion, FGM, euthanasia, and so on); however, these are causative not declarative decisions. Further, a doctor may be held to account for the decision to make the diagnosis (eg it may be the patient did not want to be diagnosed). However, in this case, the doctor is not held accountable for the content of the diagnosis but for the decision to make the diagnosis.

⁴¹ For a good overview of the issues of how to change a person’s normative situation, see David Owens, *Shaping the Normative Landscape* (2nd edn, OUP 2014). However, note that declarative decisions may change a person’s normative situation indirectly. For instance, a cancer diagnosis may mean that the patient is entitled to receive help, which means his normative situation has changed. However, it is the presence of the cancer that changed the patient’s normative information.

⁴² It may be thought that the concept of normativity is the same as the concept of morality. The two concepts are similar; however, normativity is a less demanding concept than morality. It is possible to deny that morality is a sensible concept, yet nevertheless believe in normative and ethical standards. Nothing in this chapter depends on the notion of morality, therefore, the wider concept of normativity is used.

⁴³ For instance, a doctor determining correctly that a patient has cancer is a factually right decision. To say that a prison sentence is right because it is just sentence, is a normatively right decision.

To illustrate, if a lion attacks Bob, the lion's decision changes Bob's factual situation (ie he is being attacked), but his normative situation is not changed.⁴⁴ The lion did not wrong Bob by attacking him, and neither was Bob treated unfairly or unjustly. However, if a human attacks Bob, then it is possible to ask whether it was rightful or wrongful because in this case, Bob's normative situation may have changed.⁴⁵ However, this is only possible if one deals with causative decisions. If the decision is declarative this is never possible.⁴⁶

The important implication is that, *assuming* AI cannot make normative decisions, if an AI system decides to attack you, your normative decision does not change. You can say that it was bad that it happened, but one cannot say it was rightful or wrongful.⁴⁷ This is so even if AI could 'understand' ethics perfectly and would make the ethically correct decision. To make a normative decision it is not enough for the decision-maker to make the ethically correct decision, it must be possible to hold the decision-maker to account for that decision.⁴⁸ This also explains why giving reasons is more problematic for AI judges. It is true that an AI judge can give reasons for its decisions; however, AI can only give factual, not normative, reasons. To give a 'factual reason' is to provide an explanation for how or why something happened. For instance, a doctor may say: your stomach hurts because you have eaten too many sweets. A judge can also give a factual explanation (eg I suspend your licence because you were caught drink-driving). A judge, even if it is AI, may be able to explain why the decision is ethically correct (eg people who endanger others without a good reason ought to be punished). However, this would still only be a factual explanation because all the AI system could say is that, according to some ethical standard, the sentence is correct. To give a normative explanation it is necessary that the judge can be held to account for that explanation. For declarative decisions, this is impossible but for causative decisions, this is possible because the explanation is more than just information.

This raises the question of whether AI can be held to account.⁴⁹ There is no reason why only humans can be held accountable. It seems obvious that

⁴⁴ However, Bob's normative situation may change indirectly (eg if he is attacked by a lion, other people may have a duty to provide reasonable assistance).

⁴⁵ Assuming the human has reached the age of maturity, is of sound mind, and so on.

⁴⁶ It is not suggested that something can only be normative if there is a decision-maker. It is conceivable that there are ethical systems in which it is possible to say that something is rightful/wrongful even if no one decided it. However, this does not need to concern us for the purpose of this chapter. This chapter deals with ADM. All that is claimed is that a normative decision changes the normative situation by virtue of the decision alone. It does not deal with the question if normative situations can change in other ways.

⁴⁷ In the same way that one may say that it is 'bad' to be attacked by a lion.

⁴⁸ This is broadly in line with the empirical findings by Bigman and Gray (n 17). They find that people are less averse to AI making moral decisions if AI only advises on what to do. If AI only advises, then AI makes a declarative rather than causative decision.

⁴⁹ Due to size constraints, it is not possible to provide a fully argued account. This is especially unfortunate because this question appears central to the argument provided in this chapter. However, a framework can be provided.

non-human creatures could be held accountable too (such as extra-terrestrials, angels, demons, or maybe even higher primates, squids, or other animals). There is extensive literature on the question of whether AI or algorithms can be a moral agent and some scholars hold that this is indeed possible.⁵⁰ Unfortunately, many of the arguments in this body of literature are not directly relevant for the purposes of this chapter. For instance, a prominent account that suggests that algorithms can be moral agents was put forward by Floridi and Sanders.⁵¹ Their arguments have been criticised by various scholars,⁵² and whether their arguments succeed remains a matter of debate.⁵³ The important point is that Floridi and Sanders' arguments are not directly relevant for the question of whether AI can be held accountable. They argue that being a moral agent does not mean that the agent can be held to account, and even seem to suggest that algorithms cannot be held to account.⁵⁴ While this does not mean that AI cannot be held accountable, it is striking that even prominent defenders of the view that algorithms can be moral agents do not go as far as claiming that algorithms can be held accountable.⁵⁵ Another way to approach the issue is to consider which attributes we would regard as sufficient to hold something to account. It is submitted that if something has free will, consciousness, and a concept of good and evil, then people would agree that this creature can be held to account. It may be that a creature lacking one or more of these attributes can also be held accountable. However, currently AI does not seem to have any of these attributes. Maybe in the future artificial general intelligence (AGI) will have one or all these qualities, in which case it will be necessary to revisit this issue. However, presently this seems wild speculation.

This is obviously not a fully fledged argument.⁵⁶ The argument put forward in this chapter is simply that *if* normativity is a meaningful concept (as many people believe it is), and *if* there are creatures that can be held to account and humans are one of these creatures (as many people believe), there are no good reasons to believe that AI, in its present form, can be held to account, and it follows that AI cannot make normative decisions.

⁵⁰ For an overview of the debate, see Brent Daniel Mittelstadt and others, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3(2) *Big Data & Society* <<https://doi.org/10.1177/2053951716679679>>.

⁵¹ Floridi and Sanders (n 38).

⁵² See eg Deborah G Johnson and Keith W Miller, 'Un-Making Artificial Moral Agents' (2008) 10 *Ethics and Information Technology* 123.

⁵³ As far as the present author is concerned, their argument does not work.

⁵⁴ Floridi and Sanders write, 'It can be immediately conceded that it would be ridiculous to praise or blame an AA [Artificial Agent] for its behavior or charge with moral accusation'. See Floridi and Sanders (n 38) 17.

⁵⁵ Durante (n 38) 79.

⁵⁶ For instance, one may argue that this is anthropocentric, that free will does not exist, that normativity is a meaningless concept.

6 What Decisions Should be Normative?

It may be thought that the above analysis suggests that an AI cannot be a judge because this would result in the legal system being non-normative. However, this conclusion is too quick. Imagine a system in which every car is fitted with a device that automatically deducts \$100 if the car exceeds the speed limit. In this case, no judge imposes the \$100 fine, but one can still ask whether the \$100 deduction was rightful. The reason why this is possible is because somebody (who can be held to account) decided to create the system where \$100 is deducted for speeding. Compare this to the situation where no one made that decision, but speeding drivers lose \$100 anyway (eg imagine that the running costs of a car go up by \$100 when one drives at high speeds). The factual situation of speeding drivers is exactly the same in both scenarios. However, only in the former scenario is it possible to ask whether it is fair that speeding drivers lose \$100. For the judicial system to be normative it is important that there is at least one decision-maker that can be held to account for the outcome, but not necessarily a judge. This raises the question of which decisions need to be normative for the entire system to be normative. Again, due to size constraints, it is not possible to give a fully fledged answer, but only to give some pointers.

There is no hard-and-fast rule as to how to make this determination; however, one important aspect is how much uncertainty is involved in a decision. For instance, if I know for certain that if I let my dog off the leash he will attack you, then the relevant decision for which I can be held accountable is the decision of 'letting my dog attack you'. On the other hand, if I know that there is only a small chance that my dog will attack someone, then the decision for which I can be held to account is 'taking the chance that my dog will attack you'. If it was impossible to predict that my dog will attack you, then I cannot really be held accountable for the attack. (The question of legal responsibility is obviously different; eg the law may impose strict liability.⁵⁷) For the legal system, this means the following: if a legislator could perfectly foresee every case that may arise under a specific law and provide exactly what the legal consequences are in each case, then the lawmaker could be held accountable for the outcome in each case. However, it is neither possible to foresee every case nor to write a 'perfect' law. Therefore, it is inevitable that the courts will need to make some decisions. For instance, the lawmaker may impose a tax on biscuits, but the courts need to decide whether something is a biscuit or not.⁵⁸ If the decision by the court is not normative then the normativity of the legal system decreases. If the court's decision is normative then the normativity of the

⁵⁷ For an example of the application of strict liability under English tort law, see *Rylands v Fletcher* [1868] UKHL 1.

⁵⁸ For instance, famously in a UK tax case a Value Added Tax (VAT) tribunal had to decide whether Jaffa cakes are biscuits or cakes: see *United Biscuits* [1991] LON/91/0160.

legal system increases. The extent to which normativity increases depends on how uncertain it is that something is a biscuit. If the lawmaker could specify perfectly whether something is a biscuit or not, then it does not matter whether the court's decision is normative. If there is only a little uncertainty, then the court making normative decisions increases the normativity of the legal system by only a little. If there is a lot of uncertainty then it increases the normativity of the legal system by a lot.

Of course, the same reasoning does not only apply to court decisions but to every decision within a judicial system (eg decisions by court officers, police officer, tax inspectors, and so on) This analysis explains features of some legal systems which otherwise are difficult to explain. A striking example is the way executions are conducted in some jurisdictions.⁵⁹ Some judicial systems go to great lengths to ensure that it is very difficult to hold an executioner to account for the execution. A common way of doing this is to use more than one executioner (eg firing squads).⁶⁰ This ensures that the death of the prisoner does not depend on an individual executioner's decision, making the decision *de facto* non-causative, which by implication means it is non-normative. One of the benefits of using firing squads is to remove possible feelings of guilt from an executioner. However, usually, no measures are taken to remove feelings of guilt from judges. This is remarkable because there is no reason to suppose that executioners have more fragile temperaments than judges. More importantly, even without the use of firing squads an executioner's decision contributes very little to the death of the person because if one executioner refuses to carry out the execution another executioner can usually be found easily. In that sense, the decision of an executioner is non-causative. Judges' decisions, on the other hand, usually matter a great deal. It seems, if anything, judges, rather than executioners, need to be protected from feelings of guilt. However, according to the account put forward in this chapter, this set-up makes perfect sense. There is little uncertainty around an individual executioner's decision and the decision is largely non-causative anyway. Therefore, the legal system does not lose much of its normative character, if executioners do not make normative decisions.⁶¹ Another example are juries. In countries where

⁵⁹ It may be considered somewhat strange to write about executions in this context because many people (including the author) hold that the death penalty is morally wrong. The reader is kindly asked to put this issue aside for the sake of the argument the chapter is making.

⁶⁰ A firing squad makes the decision of an executioner non-causative because even if one executioner would not have fired, the person in front of the firing squad would still have died because the other executioners would have fired. An additional method is to swap one live bullet for a blank bullet, so that there is always the possibility that the shot fired by one squad member was ineffective. In the case of an execution by electric chair or hanging, sometimes multiple executioners activate the mechanism even if one activation is sufficient to bring about the death of the condemned.

⁶¹ This further suggests why dignity-based approaches to ADM are wrong or incomplete. The use of firing squads means that the decision to kill a human is already often made by something that cannot be held to account. One can, of course, argue that the death penalty is an affront to human dignity but this has nothing to do with AI.

juries are used, often they are tasked to make declarative decisions (ie to determine whether certain facts have occurred). Judges make the normative decision about what sentence to pass (as well as instructing the jury). It is striking that the jury often operates like a black box.⁶² Usually, juries give no reasons for their decisions and the verdict may only be single word ('guilty' or 'not guilty'). According to the analysis above, this makes sense. Juries do not make normative decisions, therefore it is not a problem that juries act like entities to which normativity does not apply.

7 Framework for Which Decisions Can Be Made by AI

The above is not a fully fledged answer to the question of how to decide which decisions should be normative. However, it enables the construction of a framework to determine which decisions can be made by AI. The first question that needs to be answered is whether one is dealing with a causative or declarative decision. If it is the latter, concerns about the normativity of the decision do not apply. If the decision is causative, then it needs to be determined whether normativity plays a role. For instance, normativity may be important for the judicial system but may not play any role for an activity like cleaning. (Most people do not care whether it is right, fair, or just that their apartment is clean, they just want it clean, and in fact, most people do not have a problem with AI cleaners.) If normativity is important, then it needs to be determined whether normativity is an issue for the activity itself or for the wider system the activity is embedded in. If the latter, it needs to be determined which decisions within that system need to be normative, and this will depend on the degree of uncertainty involved in each decision. Lastly, it needs to be considered how to balance the desire for normativity with other values. As normativity is not an absolute value, and it needs to be balanced with other competing goals.

How this works may be illustrated through some examples. Consider AVs, especially the case of driverless cars and driverless trains. AVs make causative decisions; therefore, normativity may be an issue. AVs may be considered part of the wider system of transport; however, normativity is probably not an important issue for transport. Yet, normativity is important for decisions involving the risk of life or bodily harm to humans. The next issue is uncertainty. Trains operate in a more predictable environment than cars, therefore, it is less important that trains make a normative decision than cars.⁶³ Therefore, (contrary to what may be thought) there is an ethically relevant distinction between driverless trains and driverless cars. Nevertheless, even for cars, situations where it is important that the decision

⁶² For a different take on juries acting as a black box, see Chesterman (n 12) 3.3.2.

⁶³ A similar argument could be made for pilotless planes, as planes also seem to operate in a more predictable environment than cars.

is normative are probably quite rare.⁶⁴ Thus, the potential loss of normativity is small. Lastly, considering that driverless cars have the potential to save many lives (as well as other potential benefits), the small loss in normativity may not be that important.

AWS are a different matter. AWS make causative decisions, and we have a strong interest in war being normative.⁶⁵ How much uncertainty there is involved depends on the specifics of the weapon systems used. Heat-seeking missiles seem relatively predictable. However, fully autonomous killer drones that self-select their target are more uncertain. Therefore, using the latter has the potential to greatly diminish the normative character of war.⁶⁶ AWS may have certain benefits; however, they also have considerable risks.⁶⁷ How to exactly balance these benefits and risks with the loss of normativity they may bring is outside the scope of this chapter; however, on a preliminary basis, it seems that due to the loss of normativity and potential risks, AWS should not be used.

A similar analysis can be conducted for AI in the legal system. For instance, the analysis in this chapter suggests that AI lawyers providing legal advice and opinions is less problematic than having AI judges. Further, if a particular area of law is highly predictable, then it is less necessary for judges to make normative decisions, and therefore AI judges may be suitable. It could also be that in some areas of law, normativity is less important than in others.⁶⁸ Last, the value of normativity needs to be balanced with other goals. For instance, a significant problem with current legal systems is that they are often slow and expensive, and using AI judges *may* make access to justice quicker and cheaper.⁶⁹ Speed and costs are important considerations, which need to be balanced against a potential loss of normativity.

8 Conclusion

This chapter suggests that there are indeed certain decisions AI cannot make, namely normative decisions. Normative decisions are important because they allow one to say that a certain decision was rightful or wrongful. This has

⁶⁴ For instance, cases where a driver has to decide between what lives should be put at risk are probably quite rare.

⁶⁵ Recall that normative does not mean good, for the purpose of this chapter it means that it is possible to say that war is wrongful or rightful.

⁶⁶ This way of thinking may also explain why heat-seeking missiles are different from landmines. The outcome of heat-seeking missiles is highly determined but the outcome of landmines is not. With the latter there is large random element because one does not know who or what will trigger it.

⁶⁷ A benefit may be that AWS could be used for rescue missions of human soldiers. It may also be that if AWS only fight each other then war becomes less harmful to humans. The risks of AWS are that they go 'rogue' and actually increase the harm done to humans.

⁶⁸ Maybe normativity is less important in financial regulation than it is criminal law.

⁶⁹ For a broader discussion on how technology may help to improve the justice system, see eg Richard Susskind, *Online Courts and the Future of Justice* (OUP 2019).

implications for human rights. It needs to be determined whether it is permissible under human rights to change something that is normative into something that is non-normative. For instance, does the right to a fair trial mean that a trial must be normative? However, there is an even more general issue. Human rights are about protecting rights, and rights are a normative concept.⁷⁰ It was shown above that AI has the potential to turn something normative into something non-normative (to put it poetically, it may turn ‘killing’ into ‘dying’ or a ‘fine’ into ‘having less money’). This may lead to a situation where there are fewer cases to which rights apply, which in turn means fewer cases to which human rights apply. In other words, ADM may ‘hollow out’ human rights. On the other hand, AI may also bring many potential benefits, and it would be a false dichotomy to postulate that we must choose between hollowed-out human rights and the benefits of ADM .

⁷⁰ For instance, the right to life protects primarily the *right* to life rather than life itself. If a person dies from natural causes at a ripe old age, this would not ordinarily be considered a human rights violation.

33

Smart Cities, Artificial Intelligence, and Public Law

An Unchained Melody

Sofia Ranchordás

1 Introduction

This chapter discusses how smart cities—in particular through the use of artificial intelligence (AI) in the provision of public services—deepen existing inequalities between citizens and governments and how traditional public law principles, rights, and doctrines fail to correct them.¹ At the heart of this discussion is the premise that the relationship between citizens and governments is by definition unequal.²

Since the creation of modern states, government has been the sole provider of several public services; it has had the power to unilaterally shape how citizens exercise their rights and have access to social assistance; and nowadays, more than ever before, government has significant volumes of information on citizens.³ Indeed, there is a ‘natural asymmetry of information between those who govern and those whom they are supposed to serve’.⁴ However, with the digitalisation, datafication, and automation of government transactions, this asymmetry is increasing for vulnerable citizens, affecting at times their ability to exercise rights and have access to the public services they are entitled to.⁵ This asymmetry is visible at different levels

¹ More generally on smart cities and urban inequality, see Andrea Caragliu and Chiara F Del Bo, ‘Smart Cities and Urban Inequality’ (2022) 56 *Regional Studies* 1097.

² Adriaan Mallan, *Lekenbescherming in het Bestuursprocesrecht* (Wolf 2014).

³ Frank Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Society* (Harvard UP 2015); Triin Vihalemm and Margit Keller, ‘Consumers, Citizens or Citizen-consumers? Domestic Users in the Process of Estonian Electricity Market Liberalization’ (2016) 13 *Energy Research & Social Science* 38; Sofia Ranchordas, ‘Citizens as Consumers in the Data Economy: The Case of Smart Cities’ (2018) 7 *Journal of European Consumer and Market Law* 154.

⁴ Joseph E Stiglitz, ‘Transparency in Government’ in World Bank (ed), *The Right to Tell: The Role of Mass Media in Economic Development* (Word Bank 2002) 27.

⁵ Egidijus Barcevičius and others, *Exploring Digital Government Transformation in the EU: Analysis of the State of the Art and Review of Literature* (JRC Publications Office of the European Union 2019) 9; Karl Kristian Larsson, ‘Digitization or Equality: When Government Automation Covers Some, But Not All Citizens’ (2021) 38(1) *Government Information Quarterly* 101547.

of digital government, including cities where the dissemination of smartification narratives is on the rise.

The conferral or denial—either online or offline—of a licence or a permit to a citizen is not only a mere bureaucratic act but also a textbook example of a clear power asymmetry between citizens and public authorities. For citizens, a licence—which, typically, can only be issued by one public authority—may determine their ability to make a living, exercise their profession, have a hobby, and build a roof over their heads. Administrative law regulates extensively these government transactions and protects citizens *reactively* against potential abuses of power by public authorities. As government transactions shift in nature and become more predictive rather than reactive, the power asymmetry between citizens and governments is also changing.⁶

This chapter argues that traditional public law instruments may at times offer insufficient protection to vulnerable citizens in the context of smart cities. I consider not only (typical) smart urban solutions such as intelligent transportation systems, waste and water management, and smart buildings but also other automated systems used for digital enforcement, including predictive policing and risk management in welfare services. This last category of public services is not always associated with smart cities as it is an often-overlooked part of cities' tasks. However, as cities become important welfare services providers, especially in Northern Europe, they also start employing risk indicators to predict the risk of fraud (the controversial Dutch 'SyRI' algorithmic anti-fraud system is an example hereof).⁷

At first sight, it could be argued that the use of smart urban solutions to predict welfare fraud in low-income neighbourhoods does not in theory change citizens' legal situation or create a significant disadvantage.⁸ The same could be argued for the use of smart solutions to predict the fluctuations in energy use (smart grids) or the occurrence of crime (predictive policing). AI is employed in predictive policing to forecast statistically where crime is more likely to occur rather than react to it; automation is employed in welfare fraud enforcement to create risk profiles in

⁶ On the digitalisation of government transactions and the relationship between citizens and public authorities, see IDP, 'Wait No More: Citizens, Red Tape, and Digital Government' (IDP, 2018) <<http://dx.doi.org/10.18235/0001150>>.

⁷ SyRI is an acronym for the 'System Risk Indication' developed by the Dutch government to profile individuals based on large pools of personal and sensitive data that had been collected from a range of public bodies. In January 2020, the District Court of The Hague ruled that SyRI violated the right to privacy and discriminated in particular against underrepresented groups. This was one of the first cases to challenge the use of risk-scoring software. See *NJCM et al v the Netherlands* [2020] C-09-550982-HA ZA 18-388, ECLI:NL:RBDHA:2020:865

⁸ *NJCM et al v the Netherlands* [2020] C-09-550982-HA ZA 18-388, known as the 'Syri case' in which the algorithmic anti-fraud tool was found to be unlawful and contrary to the European Convention on Human Rights (ECHR). See also Valery Gantchev, 'Data Protection in the Age of Welfare Conditionality: Respect for Basic Rights or a Race to the Bottom?' (2019) 21(1) European Journal of Social Security 3; Marvin van Bekkum and Frederik Zuiderveen Borgesius, 'Digital Welfare Fraud Detection and the Dutch SyRI Judgment' (2021) 23(4) European Journal of Social Security 323.

order to narrow down the number of individuals that should be investigated for potential fraud, and not to sanction them immediately.⁹ At the resemblance of many other smart urban solutions, these systems are primarily employed to optimise the allocation of local resources and are thus presented as mere policy choices.¹⁰ Nonetheless, in practice, they have far-reaching effects. For example, for citizens, additional surveillance translated into more police patrolling their neighbourhood may amount to greater stigmatisation.¹¹ For low-income groups and ethnic minorities who tend to be common targets of predictive smart urban solutions, the existence of risk-scoring systems immediately changes how they live and interact with government.¹² It adds to existing fears of bureaucracy and the phenomenon of government anxiety experienced by many citizens.¹³

Equality and inclusiveness considerations in smart cities and urban law are relatively recent discourses.¹⁴ These discussions are nonetheless crucial for several reasons. First, as more than half of the world's population lives nowadays in urban centres, cities have been called to assume an increasingly important role in crucial discussions such as migration, climate change, and the digital transformation. Second, beyond data protection legal frameworks, citizens may have more limited (legal) options to react against predictive solutions or policy decisions concerning the establishment of smart cities employing AI. Third, smart cities put in place narratives that exclude vulnerable citizens who do not fit into their vision of corporate urban spaces driven by technology. Fourth, there is a strong dissonance between the predictive nature of smart urban solutions fuelled by AI and the existing public law principles and instruments that aim to protect citizens against them primarily

⁹ Lyria Bennett Moses and Janet Chan, 'Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability' (2018) 28 (7) *Policing and Society* 806. See also Marion Oswald and others, 'Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and "Experimental" Proportionality' (2018) 27(2) *Information & Communications Technology Law* 223; P Jeffrey Brantingham, Matthew Valasik, and George Mohler, 'Does Predictive Policing Lead to Biased Arrests? Results from a Randomized Controlled Trial' (2018) 5(1) *Statistics and Public Policy* 1; Rikke Frank Jørgensen, 'Data and Rights in the Digital Welfare State: The Case of Denmark' (2021) 26(4) *Information Communication and Society* 1–16; Anne Kaun, 'Suing the Algorithm: The Mundanization of Automated Decision-Making in Public Services through Litigation' (2021) 25(14) *Information Communication and Society* 2046.

¹⁰ Andrew Guthrie Ferguson, 'Predictive Policing and Reasonable Suspicion' (2012) 62 *Emory Law Journal* 259.

¹¹ Simon Egbert and Monique Mann, 'Discrimination in Predictive Policing: The (Dangerous) Myth of Impartiality and the Need for STS Analysis' in A Završnik and V Badalič (eds), *Automating Crime Prevention, Surveillance, and Military Operations* (Springer 2021) 2.

¹² See UN Press Release, 'Landmark Ruling by Dutch Court Stops Government Attempts to Spy on the Poor—UN Expert' (UN, 2020) <www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?LangID=E&NewsID=25522>.

¹³ Herbert Kaufman, 'Fear of Bureaucracy: A Raging Pandemic' (1981) 41(1) *Public Administration Review* 1.

¹⁴ See Homi Kharas and Jaana Remes, 'Can Smart Cities Be Equitable?' (*Mckinsey*, 21 June 2018) <www.mckinsey.com/mgi/overview/in-the-news/can-smart-cities-be-equitable>; Jiska Engelbert, Liesbet van Zoonen, and Fadi Hirzalla, 'Excluding Citizens from the European Smart City: The Discourse Practices of Pursuing and Granting Smartness' (2019) 142 *Technological Forecasting and Social Change* 347.

in reactive or responsive ways. This dissonance contributes to the deepening of existing inequalities between government and citizens. Recent legal scholarship has discussed the discriminatory effect of automation; how the use of algorithms affects procedural rights, equality of arms, and the right to a fair trial; and how digital welfare fraud enforcement tends to target the most vulnerable members of our society.¹⁵ AI poses indeed multiple risks to the exercise of fundamental rights and it exacerbates historical inequalities.¹⁶ Nevertheless, this strand of legal scholarship has overlooked a key element underlying these legal problems: AI systems and the datafication of the public sector, that is, the practice of quantifying government-citizens transactions, have a fundamentally different rationale from the traditional tools that public law has at its service to protect citizens against the abuses from government.¹⁷

In simple terms, and risking a certain degree of technical and legal inaccuracy, automation in government (including in smart cities) is used to advance effectiveness and efficiency, reduce unwanted ‘noise’ (eg mistakes, inconsistencies, delayed or omitted responses to citizens’ requests), predict needs for better resource allocation, and anticipate problems before they occur.¹⁸ AI systems in government have a pre-emptive logic: they identify, ‘flag’, predict, and—possibly—prevent risks.¹⁹

¹⁵ See Danielle Keats Citron, ‘Technological Due Process’ (2008) 85(6) Washington University Law Review 1249; Solon Barocas and Andrew D Selbst, ‘Big Data’s Disparate Impact’ (2016) 104 California Law Review 671; Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin’s Press 2018); Niels Jak and Steven Bastiaans, ‘De Betekenis van de AVG voor Geautomatiseerde Besluitvorming door de Overheid: Een Black Box voor een Black Box?’ (2018) 40 Nederlands Juristenblad 3018; Alessandro Mantelero, ‘AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment’ (2018) 34(4) Computer Law & Security Review 754; Jennifer Cobbe, ‘Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decisionmaking’ (2019) 39 Legal Studies 636; Janneke Gerards, ‘The Fundamental Rights Challenges of Algorithms’ (2019) 37(3) Netherlands Quarterly of Human Rights 205; Sonja Bekker, ‘Fundamental Rights in Digital Welfare States: The Case of SyRi in the Netherlands’ (2019) 50 Netherlands Yearbook of International Law 289–307; Johan Wolswinkel, *Willekeur of Algoritme? Laveren tussen Analoog en Digitaal Bestuursrecht* (Tilburg UP 2020); Joe Tomlinson, Jack Maxwell, and Alice Welsh, ‘Discrimination in Digital Immigration Status’ (2022) 42(2) Legal Studies 315.

¹⁶ Sandra Wachter, Brent Mittelstadt, and Chris Russell, ‘Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law’ (2021) 123 West Virginia Law Review 735.

¹⁷ On the phenomenon of datafication, see Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work and Think* (John Murray 2013); Lina Dencik and Anne Kaun, ‘Datafication and the Welfare State’ (2020) 1(1) Global Perspectives 12912; Heather Broomfield and Lisa Reutter, ‘In Search of the Citizen in the Datafication of Public Administration’ (2022) 9(1) Big Data & Society <<https://doi.org/10.1177/20539517221089302>>; Lisa Reutter, ‘Constraining Context: Situating Datafication in Public Administration’ (2022) 24(4) New Media & Society 903.

¹⁸ Cary Coglianese and David Lehr, ‘Regulating by Robot: Administrative Decision Making in the Machine-Learning Era’ (2017) 105 Georgetown Law Journal 1147; David Lehr and Paul Ohm, ‘Playing with the Data: What Legal Scholars Should Learn about Machine Learning’ (2017) 51 UC Davis Law Review 653; Daniel Kahneman, Olivier Sibony, and Cass Sunstein, *Noise: A Flaw in Human Judgment* (Little Brown Spark 2021); Cary Coglianese and A Lai, ‘Algorithm vs. Algorithm’ (2022) 71 Duke Law Journal 1281; Cass Sunstein, ‘Governing by Algorithm? No Noise and (Potentially) Less Bias’ (2022) 71 Duke Law Journal 1175.

¹⁹ Law enforcement relies on a mix of reactive and proactive methods, predictive policing—increasingly used in smart cities—places its focus on the proactive or preventive dimension, see Daniel

Public law, on the contrary, provides citizens mainly with reactive and balancing tools and frameworks to correct and sanction abuses of power (eg principle of proportionality).²⁰ These tools are needed because public authorities can unilaterally determine the legal sphere of citizens. Public law has sought to compensate for underlying asymmetries between individuals and governments through institutions such as the Ombudsman, procedural mechanisms favourable to citizens (eg the transfer of proceedings to the competent court in case of mistake without detriment to applicant), principles of good administration, and more specific national administrative law principles. In the age of AI and smartification of government and public urban services, this does not suffice.

This chapter focuses on smart cities because urban centres are typically the closest locus of interaction between citizens and governments. It takes smart cities as an illustration of the dissonance between the logic of smart urban solutions and the modus operandi of public law. This chapter contributes to the literature by explaining the paradox between the way AI systems are developed and implemented and how public law protects citizens against them. It does not consider directly equal treatment and non-discrimination issues among citizens. Instead, it reflects upon how smart urban solutions reinforce the power asymmetry between governments and individuals.

This chapter is organised as follows. Section 2 discusses the use of AI systems in smart cities, providing a brief overview of smart urban solutions. Section 3 explains how certain smart urban solutions may expose or deepen existing inequalities. Section 4 discusses the disconnection between public law and smart urban solutions. Section 5 concludes with some reflections on how to conceptualise and address inequality in smart cities.

2 Smart Urban Solutions

2.1 Definition of Smart City

There is no single definition of smart city.²¹ Smart cities are commonly defined as urban centres ‘that monitor and integrate conditions of all of [their] critical

Susser, ‘Predictive Policing and the Ethics of Preemption’ in Ben Jones and Eduardo Mendieta (eds), *The Ethics of Policing: New Perspectives on Law Enforcement* (New York UP 2021) 268.

²⁰ Julian Rivers, ‘Proportionality and Discretion in International and European Law’ in Nicholas Tsagourias (ed), *Transnational Constitutionalism: International and European Perspectives* (CUP 2010); Aharon Barak, *Proportionality: Constitutionality Rights and their Limitations* (CUP 2012) 175. On proportionality and equality, see Guy Lurie, ‘Proportionality and the Right to Equality’ (2020) 21(2) German Law Journal 174.

²¹ Vito Albino, Umberto Berardi, and Rosa Maria Dangelico, ‘Smart Cities: Definitions, Dimensions, Performance, and Initiatives’ (2015) 22(1) Journal of Urban Technology 3.

infrastructures, including roads, bridges, tunnels, rails, subways, airports, seaports, communications, water, power, even major buildings, [to] better organize [their] resources, plan [their] preventive maintenance activities, and monitor security aspects while maximizing services to [their] citizens.²² Other scholarly definitions focus on smart cities' creativity, their ability to 'motivate its inhabitants and flourish in their own lives'.²³ Smart cities have heterogeneous profiles and different cities will tend to have different goals.²⁴ There is, therefore, not one model of smart city but many clusters of smart cities, each with its own focus, strategy, and approach.²⁵

While some smart cities may harness digital technology to promote sustainability through the use of smart waste and water management systems; others rely on technology primarily for law enforcement.²⁶ In this chapter, the smart city is defined as an urban centre that harnesses digital technology—including AI—to promote economic growth and optimise its public services through public and private partnerships.²⁷ This chapter acknowledges that the aimed improvement of public services may have important costs for many citizens (eg surveillance, reduced autonomy) and will not always serve them equally well. This happens because the smart city is a complex reality: a smart city is at the same time a *strategy* to ameliorate urban centres, a *technological product*, a *narrative*, and a *process*.

2.2 Smart Cities as Strategies, Products, Narratives, and Processes

As a *strategy* to urban and economic development, the smart city can be developed to foster innovation and improve the liveability of neighbourhoods and the well-being of their residents.²⁸ This strategy relies on technology to promote sustainable growth, better use of resources for saving, for example, energy or water. It employs the Internet of Things (IoT) and big data to develop data-driven policy and monitor risks (eg risks of floods), encourage connected communities through

²² Rudolf Giffinger and others, *Smart Cities: Ranking of European Medium-Sized Cities* (Vienna University of Technology 2007).

²³ Justin O'Connor and Kate Shaw, 'What Next for the Creative City' (2014) 5(3) Culture and Society 165.

²⁴ Renata Paola Dameri, 'Smart City Definition, Goals and Performance' in Renata Paola Dameri (ed), *Smart City Implementation* (Progress in IS Springer Cham 2017).

²⁵ Simon Joss and others, 'The Smart City as Global Discourse: Storylines and Critical Junctures Across 27 Cities' (2019) 26(1) Journal of Urban Technology 3.

²⁶ Sofia Ranchordas, 'Nudging Citizens through Technology in Smart Cities' (2020) 34(3) International Review of Law, Computers & Technology 254.

²⁷ Astrid Voorwinden, 'The Privatised City: Technology and Public-Private Partnerships in the Smart City' (2021) 13(3) Law, Innovation, and Technology 439.

²⁸ Andrea Caragliu, Chiara Del Bo, and Peter Nijkamp, 'Smart Cities in Europe' (2011) 18(2) Journal of Urban Technology 65; Soumaya Ben Letaifa, 'How to Strategize Smart Cities: Revealing the SMART Model' (2015) 68(7) Journal of Business Research 1414.

the construction of smart buildings, and pre-empt crime by tapping into different sources of data.²⁹ This strategic dimension is reflected in advancements at the level of smart governance, smart education, smart energy, smart buildings, smart mobility, and smart infrastructures.³⁰ A smart city is a strategy that has an important impact on different urban infrastructures, ranging from water, energy to mobility.

As a *product*, a smart city is a combination of technologies that facilitate the collection and processing of data and the transformation of cities. It is a product that is sold to city officials and often consists of different IoT packages (eg different types of sensors to measure noise, pollution, control crowds). Smart city solutions are embedded in traditional infrastructures to optimise existing services.³¹ Through different sensors placed on urban furniture (such as garbage bins or traffic lights), smart city products collect information on the city conditions (weather, pollution, water levels), its residents, commuters, and visitors, generating information on how to best allocate resources.³² AI is employed in smart cities to advance sustainability (eg analyse and optimise energy use, detect CO₂ emission levels), improve mobility (eg reduce traffic congestion), assist tourists navigate the urban maze (such as chatbots), and control crowds.³³

As a *narrative*, the smart city advertises the story of cities where citizens can move swiftly between different spaces, where tensions and delays are minimised, and individuals can have access to better public services, engaging with the latest technology, often in a sustainable way.³⁴ This narrative rarely contemplates differences between citizens and the fact that not all citizens may want or be able to engage with smart urban solutions on equal terms. The key idea of smart cities is thus ‘smartness’ and that is by itself, an all-encompassing term that while avoiding ‘metaphysical debates’, is an empty vessel that allows cities to be reconfigured without asking crucial questions: Smart for whom? At what cost? Who is smart enough to engage with smart cities? Who is left behind in the smart narrative?³⁵

The origins of the term ‘smart technology’ refer to perceptions that are closely related to smart urban solutions, namely self-monitoring analysis and reporting. ‘Smart’ is not a technical term that captures any type of use of AI or any other

²⁹ Deloitte, ‘Define your Smart City Strategy: Dive Deeper into the Six Domains of Smart Cities’ <www2.deloitte.com/us/en/pages/consulting/solutions/smart-cities-strategies.html>.

³⁰ Matthew Jewell, ‘Contesting the Decision: Living in (and Living with) the Smart City’ (2018) 32(2) International Review of Law, Computers & Technology 210, 212–13.

³¹ For a thorough analysis of the definition of smart city, see Albino, Berardi, and Dangelico (n 21).

³² Ryan Calo and others, ‘Push, Pull, and Spill: A Transdisciplinary Case Study in Municipal Open Government’ (2015) 30 Berkeley Technology Law Journal 1900, 1902. See also Peter A Johnson, ‘Models of Direct Editing of Government Spatial Data: Challenges and Constraints to the Acceptance of Contributed Data’ (2017) 44(2) Cartography and Geographic Information Science 128.

³³ Ryan Mark, ‘Ethics of Public Use of AI and Big Data: The Case of Amsterdam’s Crowdedness Project’ (2019) 2(2) The ORBIT Journal 1.

³⁴ Lucas Melgaço and Rosamunde van Brakel, ‘Smart Cities as Surveillance Theatre’ (2021) 19(2) Surveillance & Society 244.

³⁵ Mireille Hildebrandt, ‘Smart Technologies’ (2020) 9(4) Internet Policy Review 1.

technology, even though the use of AI in smart urban solutions is mounting. As Hildebrandt speculates, '[smart] may have been initiated as a marketing term, hoping to lure investors and potential users'.³⁶ The marketing aspect of the smart city has been exposed in specialised literature which has argued that smart cities primarily reflect corporate interests rather than the real needs of citizens.³⁷

Finally, a smart city is a *process*. While there are cities such as Songdo in South Korea which have been designed from scratch as smart cities, most urban centres are in the process of becoming smart through pilot projects and other experiments (such as living labs) which seek to reconfigure public spaces and public services.³⁸

Smart cities are, nonetheless, strategies, products, narratives, and processes of exclusion and extraction for the citizens who do not fit in their vision. The following section discusses how smart cities and their smart solutions—increasingly powered by AI—can exclude vulnerable citizens (eg individuals with low literacy, senior citizens).

3 Smart Cities and Exclusion

In theory, the use of digital technology to provide public services in the context of smart cities should help close the gap between citizens and governments. In practice, smart cities harness different technologies to improve the liveability of urban centres, advance economic growth, approximate citizens from local governments, and have the potential to turn the city into spaces of equal participation. AI enables many of the smart urban solutions that are susceptible of facilitating the contact between citizens and local public authorities with real-time information, reduced response time to queries, and more responsive solutions to urban problems.

AI and other technologies used in smart cities are not only part of the product offered to city officials and citizens; they are rights intermediaries that transform socio-spatial relations, extract, and exclude at several levels.³⁹ First, it is unclear who the citizen of a smart city is or should be. Cities are designed for its residents, commuters, and visitors. The idea of citizen-centred smart cities promoted by the tech companies behind many smart urban solutions has been criticised on multiple accounts. To illustrate, it is unclear what citizens smart cities have in mind (eg tech-savvy citizens or senior citizens?) and how to involve citizens who do not fit

³⁶ ibid. See also Mireille Hildebrandt, *Smart Technologies and the End(s) of Law* (Edward Elgar 2015).

³⁷ Robert G Hollands, 'Will the Real Smart City Stand Up?' (2008) 12(3) City 302; Mark Deakin and Husam Al Waer, 'From Intelligent to Smart Cities' (2011) 3(3) Intelligent Buildings International 140; Ola Öderström, Till Paasche, and Francisco Klauser, 'Smart Cities as Corporate Storytelling' (2014) 18(3) City 307.

³⁸ Bo Wang, Becky PY Loo, and Gengzhi Huang, 'Becoming Smarter through Smart City Pilot Projects: Experiences and Lessons from China since 2013' (2021) 29(4) Journal of Urban Technology 3.

³⁹ James Ash, Rob Kitchin, and Agnieszka Leszczynski, 'Digital Turn, Digital Geographies?' (2018) 42(1) Progress in Human Geography 25.

within technological narratives.⁴⁰ Because of this blind spot, smart cities reinforce distributional justice problems as they mainly benefit technical elites without accounting for the specific gender and racial needs.⁴¹ Research on India's smart cities has demonstrated how women are repeatedly excluded from this narrative.⁴² In Ahmedabad, a smart city under development, thousands of informal recyclers—often women of the Dalit caste who are traditionally discriminated against and regarded as 'untouchable'—have seen their livelihoods affected by the use of smart city technologies. The use of smart waste management meant that waste collection started being done around the clock—mostly at night—when it was not safe for women to travel to landfills.⁴³ Also, in Latin America, there is evidence that smart urban solutions on law enforcement are disproportionately used in low-income neighbourhoods, including for crimes such as domestic violence which also occur in high-income neighbourhoods.⁴⁴

Second, 'smart citizenship' has been viewed as a passive reality where citizens are depicted as consumers of smart urban solutions.⁴⁵ This neoliberal vision of citizens as consumers has also been built upon the idea of 'choice', assuming that smart cities are markets where citizens can choose products and services and where they can be nudged by smart urban solutions to make better choices.⁴⁶ This is far from the reality where cities are poorly funded, citizens are not individuals with the choice to walk away from the city where they live as they would in a market situation, and their entitlements are not based on choice but on constitutional or legal rights to public services. Last, smart cities have become highly privatised spaces where city officials become increasingly dependent on private technology companies such as Google, generating a practical effect of privatisation and outsourcing of public services to private companies.⁴⁷

⁴⁰ Martijn de Waal and Marloes Dignum, 'The Citizen in the Smart City. How the Smart City Could Transform Citizenship' (2017) 59(6) *Information Technology* 263.

⁴¹ Marit Rosol and Gwendolyn Blue, 'From the Smart City to Urban Justice in a Digital Age' (2022) 26(4) *City* 684.

⁴² Ayona Datta, 'The "Smart Safe City": Gendered Time, Speed, and Violence in the Margins of India's Urban Age' (2020) 110(5) *Annals of the American Association of Geographers* 1318.

⁴³ J Wittmer, 'Smart City Technologies Pose Serious Threats to Women Waste Workers in India' (*The Conversation*, 19 May 2022) <<https://theconversation.com/smart-city-technologies-pose-serious-threats-to-women-waste-workers-in-india-182365>>.

⁴⁴ Beatriz Botero Arcila, 'Latin American Cities in the Fourth Industrial Revolution: The Potential and Social Risks of Smart-Cities Technologies' (2019) 42 *Latin American Policy Journal* <<https://ssrn.com/abstract=3653493>>.

⁴⁵ Paolo Cardullo and Rob Kitchin, 'Being a "Citizen" in the Smart City: Up and Down the Scaffold of Smart Citizen Participation in Dublin, Ireland' (2019) 84(1) *GeoJournal* 1. See also Robert Cowley, Simon Joss, and Youri Dayot, 'The Smart City and Its Publics: Insights from Across Six Cities' (2018) 11(1) *Urban Research and Practice* 53.

⁴⁶ Alberto Vanolo, 'Cities and the Politics of Gamification' (2018) 74 *Cities* 320. For a critique of this vision and a review of the literature, see Paolo Cardullo and Rob Kitchin, 'Smart Urbanism and Smart Citizenship: The Neoliberal Logic of "Citizen-Focused" Smart Cities in Europe' (2019) 37(5) *Environment and Planning C: Politics and Space* 813.

⁴⁷ Evgeny Morozov and Francesca Bria, 'Rethinking the Smart City. Democratizing Urban Technology' (*City Series* #5, 2018) <www.rosalux-nyc.org/rethinking-the-smart-city/>.

Third, smart urban solutions generate new security and social vulnerabilities due to weak software security, poor data encryption, use of insecure legacy systems, lack of maintenance, limited training of civil servants in cities, and disproportionate burdening of low-income citizens.⁴⁸ The surveillance generated by smart urban solution may produce more harm than benefits. The limited effectiveness of predictive policing has been an example of a smart urban solution that has promised more than it has delivered. Empirical scholarship has demonstrated that current predictive policing techniques do not evidence a real decrease in crime.⁴⁹

In conclusion, smart urban solutions have in theory the potential to improve cities and the lives of their residents. However, smart cities are corporate narratives without an inclusive vision of citizens that often disregard local customs, needs, and digital capabilities of under-represented groups, thus creating social enclaves.⁵⁰ As Beatriz Botero Arcila argues, smart city partnerships with private technology companies are connected to historical perspectives of urban economic development which did not revolve around issues of inequality.⁵¹ Historically, cities have been denied either the powers to act against inequality or have had fewer incentives to do so given the focus on central government social welfare policies.⁵² However, the situation is changing. Cities' powers have increasingly acquired international visibility and strength, thus turning cities into global actors in the advancement of sustainability and human rights protection.⁵³ In addition, in many European countries, we observe that there is a growing decentralisation of social welfare to the local level which is not yet mirrored in existing smart city narratives.

AI is employed in the development of different smart urban solutions such as intelligent transportation systems, smart grids, and waste management. Many of the risks associated with the employment of AI systems in government are also present in the context of smart cities. Nevertheless, in the context of smart cities, AI can further reinforce the degradation of urban problems due to the imposition of top-down, corporative narratives that assume that all citizens are benefitted by smart cities, the close levels of surveillance, and the limited ability of citizens to opt-out of these systems. Digital exclusion raises the question of whether citizens

⁴⁸ Rob Kitchin and Martin Dodge, 'The (In)Security of Smart Cities: Vulnerabilities, Risks, Mitigation, and Prevention' (2019) 26(2) *Journal of Urban Technology* 47.

⁴⁹ Albert Meijer and Martijn Wessels, 'Predictive Policing: Review of Benefits and Drawbacks' (2019) 42(12) *International Journal of Public Administration* 1031.

⁵⁰ Katherine Harrison, 'Who Is the Assumed User in the Smart City?' in Vangelis Angelakis and others (eds), *Designing, Developing, and Facilitating Smart Cities: Urban Design to IoT Solutions* (Springer 2017) 17.

⁵¹ Beatriz Botero Arcila, 'The Place of Local Government Law in the Urban Digital Age' (SSRN, 17 May 2021) <<https://ssrn.com/abstract=3848202>>.

⁵² Richard Schragger, *City Power: Urban Governance in a Global Age* (OUP 2016).

⁵³ Yishai Blank, 'Localism in the New Global Legal Order' (2007) 47 *Harvard Journal of International Law* 263; Barbara Oomen, 'Human Rights Cities: The Politics of Bringing Human Rights Home to the Local Level' in J Handmaker and K Arts (eds), *Mobilising International Law for 'Global Justice'* (CUP 2020) 208.

could resort to law to exercise their rights. The following section discusses why the narrative and operational logic of AI systems used in smart cities are in dissonance with existing public law principles and rules that should protect citizens against digital exclusion.

4 AI in Smart Cities: Three Paradoxes

In this section and in this chapter, I set aside discussions on data protection and privacy law which may confer additional protection to citizens in surveillance contexts in smart cities.⁵⁴ This section analyses the mismatch between the predictive, do-it-yourself, and systematic rationales of smart urban solutions and the reactive—or at most—responsive nature of public law. I consider here the logic underlying local or urban law and administrative law from a theoretical perspective. Different jurisdictions may have different instruments in place. This section offers a first reflection on the dissonance between the way smart urban solutions are designed and implemented and the way rights and legal principles are designed to protect citizens against it. I focus on three aspects of the design and operation of smart urban solutions that are at crossroads with the rationale of public law: (i) the predictive nature of smart urban solutions; (ii) the design of digital public services with ‘average citizens’ in mind; and (iii) their systemic nature and potential failures.

4.1 Prediction versus Public Law

Smart urban solutions (eg intelligent transport systems, smart grids, predictive policing) increasingly operate in *predictive* terms.⁵⁵ As explained earlier, the use of predictive systems ranges from predicting energy needs, number of commuters during rush hour to anticipating crime areas, welfare fraud, and homelessness. The goal of these predictions is to better allocate public resources. When predictive systems used in smart cities stigmatise—either directly or indirectly—communities that are already underrepresented, these citizens may have limited legal protection against these outcomes because local and administrative law—often studied together in civil law countries—are not designed to operate on predictive terms.

As most fields of law, administrative law is reactive and at the very best, *responsive* to changes. In many civil law systems, in the context of administrative law, citizens gain access to judicial protection when their legal sphere is altered by an

⁵⁴ Lilian Edwards, ‘Privacy, Security and Data Protection in Smart Cities: A Critical EU Law Perspective’ (2016) 2(1) European Data Protection Law Review 28.

⁵⁵ See Christopher Grant Kirwan and Fu Zhiyong, *Smart Cities and Artificial Intelligence* (Elsevier 2020).

administrative decision or act. For example, a citizen submits a statement of objection following a fine because a smart bin detected that they did not comply with local recycling rules. Most smart urban solutions are not employed to determine sanctions immediately but to predict needs. They pre-empt rather than sanction. Legal protection against unfavourable decisions does not address pre-emption.

The anticipatory nature of law has been discussed in the literature but existing debates have remained thus far at a theoretical level.⁵⁶ In the specific case of social welfare fraud, the law in the books teaches us that existing rules are not meant to sanction before there is evidence of fraud but predictive systems can be the first step to a certain sanction for ethnic minorities and other underrepresented groups.⁵⁷ However, administrative law does not offer many legal mechanisms to defend citizens against predictive actions unless they take the form of unwarranted profiling or grave privacy violations.

Thus far, administrative law has not offered satisfactory legal answers to the dissonance under analysis because it was traditionally conceived to constrain legal discretion, avoid abuses, issue rules and guidance that define the legal position of citizens in a reactive rather than anticipatory way, and adjudicate conflicts between citizens and government.⁵⁸ Administrative law organises public administration and protects citizens from the state in *a reactive rather than predictive way*. The principles of good administration and other standards guide ex ante the processes to be followed by public authorities (eg transparency, timeliness of decision-making) but do not address many of the policy choices underlying smart urban solutions.⁵⁹

Since predictive AI will continue to be used in the public sector within the constraints of the AI Act, it is important to think about how we can reshape administrative law to protect citizens more effectively before predictive actions from public authorities.

4.2 Do-It-Yourself Public Services

Automation in the public sector—including in smart cities—assumes that citizens can engage with technology without human assistance. Unless citizens have

⁵⁶ Sofia Ranchordas and Mattis van 't Schip, 'Future-Proofing Legislation for the Digital Age' in Sofia Ranchordas and Yaniv Roznai (eds), *Time, Law, and Change: An Interdisciplinary Study* (Hart 2020) 347.

⁵⁷ Sofia Ranchordas and Ymre Schuurmans, 'Outsourcing the Welfare State: The Role of Private Actors in Welfare Fraud Investigations' (2020) 7 European Journal of Comparative Law and Governance 5; R De la Feria, 'Tax Fraud and Selective Law Enforcement' (2020) 47 Journal of Law and Society 266.

⁵⁸ Richard B Stewart, 'Administrative Law in the Twenty-First Century' (2003) 78 New York University Law Review 437.

⁵⁹ Arjan Widlak, Marlies van Eck, and Rik Peeters, 'Towards Principles of Good Digital Administration' in M Schuilenberg and R Peeters (eds), *The Algorithmic Society: Technology, Power, and Knowledge* (Routledge 2020) 67.

documented disabilities and are entitled to assistance, most citizens are ‘on their own’ and are all treated as ‘average citizens’.⁶⁰ However, there is no such thing as an ‘average citizen’ as research shows that everyone can be vulnerable at some point in life.⁶¹ Also, understanding or knowing one’s rights does not always mean being able to exercise them. For example, about 20 per cent of the population in the Netherlands—a highly connected country—has low literacy and finds it difficult to engage with (digital) bureaucracy.⁶² As a result, in the field of social welfare which is nowadays partly managed at local level in many European countries, many citizens in need do not apply for the benefits they are entitled to.⁶³ Others who do, often make mistakes when filling in digital forms in an attempt to exercise their rights.⁶⁴

Unless citizens have documented disabilities, current administrative law systems do not offer additional assistance.⁶⁵ This leaves vulnerable citizens in a challenging position because when they make honest mistakes because of lack of skills or knowledge, AI systems may often not distinguish between honest mistakes and fraud. The personal situation of citizens (eg low literacy, sudden unemployment, or financial changes susceptible of affecting the citizen’s mental capacity) will rarely be considered by AI systems which are subtracted of the human element. AI in the public sector is thus designed for average citizens who can rely on a do-it-yourself administration. For all the others without a documented disability, traditional administrative law mechanisms do not offer meaningful solutions.

Vulnerable citizens (eg low literacy, senior citizens, low income) do not experience the traditional asymmetry vis-à-vis government the same way as the tech-savvy citizens that fit in the smart city narrative. Extant legal principles and instruments do not consider this additional layer of inequality. For low-income citizens, the gap between them and government often feels deeper as they have more contact with public authorities (eg social welfare benefits) and thus experience administrative rules and procedures on a regular basis as more onerous or

⁶⁰ Here, I refer to the average citizen in the public law realm as the individual who can exercise rights without assistance. The ‘average citizen’ can also be regarded as an economic concept or demographic category.

⁶¹ Nina A Kohn, ‘Vulnerability Theory and the Role of Government’ (2014) 26(1) *Yale Journal of Law and Feminism* 1; Martha A Fineman, ‘The Vulnerable Subject: Anchoring Equality in the Human Condition’ (2008) 20 *Yale Journal of Law and Feminism* 9; Anne-Greet Keizer, Wil Tiemeijer, and Mark Bovens, *Why Knowing What to Do is Not Enough: A Realistic Perspective on Self-Reliance* (WRR 2019) 7, 29.

⁶² Stichting Lezen en Schrijven, *Laaggeletterden: Achterblijvers in de Digitale Wereld* (Expertisecentrum Beroepsonderwijs 2015) <www.lezenenschrijven.nl/uploads/editor/ecbo.15-217-Laaggeletterden-achterblijvers-in-de-digitale-wereld-web.pdf>.

⁶³ Keizer, Tiemeijer, and Bovens (n 61).

⁶⁴ Sofia Ranchordas and Luisa Scarcella, ‘Automated Government for Vulnerable Citizens’ (2022) 72 *William & Mary Bill of Rights Journal* 373.

⁶⁵ For a critique, see Katherine Macfarlane, ‘Disability without Documentation’ (2021) 90 *Fordham Law Review* 60.

emotionally taxing than the ‘average citizen’.⁶⁶ As the following subsection explains, these citizens are also more often victims of system errors as they fit in categories captured by risk indicators.

4.3 Systems versus Individual Human Flaws

In many jurisdictions, administrative law acknowledges the need to correct the asymmetry between citizens and governments. It does so through principles and instruments that aim to constrain discretion, protect citizens against abuses, and assist citizens in government transactions. An example of this is the principle for the compensation of inequality in Dutch administrative law which requires public authorities and administrative judges to take active measures to ensure that citizens are compensated for this inequality.

In addition, administrative law addresses individual human flaws rather than system errors. The employment of AI in the public sector raises nonetheless issues that are not only related to individual human flaws but that are rather the result of system failures (inaccurate systems, broader historical biases in the system). In the specific case of smart cities, the narrative underlying this corporate story is also a systematic element to consider.

In continental systems of public law, especially those inspired by the German legal tradition, administrative law is a fairly recent field of law which was designed specifically to address human flaws (eg abuse of discretion to favour friends or family with public funds) in human-to-human relations. At stake is a human (the citizen) who needs assistance and a human (civil servant representing a public authority) who needs to be regulated. Administrative law considers all relevant circumstances to the provision of public services, the allocation of public money, and the use of discretion. However, in the context of digital public services with a predictive character, it may be difficult to apply existing administrative law rules. The debate regarding the use of predictive systems is often dismissed by arguing that there should always be ‘a human in the loop’ in the case of automated decision-making. However, smart urban solutions rarely make traditional decisions in the sense of a final determination of the legal sphere of citizens’ acts (regulation or adjudication). Also, we know that in social welfare fraud enforcement cases, humans tend to trust the information that is produced by AI systems.⁶⁷ AI does not only subtract humans from the digital administrative state. It also has the potential to dehumanise if no adequate measures are taken. Its predictive character, associated

⁶⁶ Julian Christensen and others, ‘Human Capital and Administrative Burden: The Role of Cognitive Resources in Citizen-State Interactions’ (2020) 80 Public Administration Review 127; see also George Frederickson, *Social Equity and Public Administration: Origins, Developments, and Applications* (Routledge 2010).

⁶⁷ Ranchordas and Schuurmans (n 57).

with the lack of assistance, and its dehumanising character may leave vulnerable citizens at the mercy of smart urban solutions and their optimisation narratives.

5 Conclusion

AI, IoT, and biometrics in smart cities have deepened existing inequalities, biases, and exclusion. AI, in particular, is employed to predict crime and fraud before they occur, identify protesters, and control crowds.⁶⁸ This chapter has argued that the use of AI in smart cities is particularly problematic because it deepens the asymmetry that traditionally characterises government transactions. This asymmetry which explains the existence of administrative law, its principles, and rules, is not adequately addressed in the digital context. Instead, there is a fundamental dissonance between the rationale of smart urban solutions and the principles and instruments employed by public law to close the inequality gap between citizens and government.

While most AI systems applied in smart cities are predictive as they seek to prevent, for example, crime, traffic congestion, or littering, public law principles and tools (eg principles of good administration) remain primarily reactive or responsive, at best. In addition, public law was designed with human and individual flaws in mind, not systemic errors and datafication which place citizens in novel categories. Public law also does not accommodate the smart city as a corporate narrative which, with its attempts to optimise citizenship, inevitably excludes thousands of citizens. This chapter offered a first exploration of this problem. Administrative law needs to look above and beyond human flaws and constructs such as discretion. The novel challenges of administrative and local law, particularly in cities, are more complex and cannot be reduced to reactive or responsive instruments. Citizens are nowadays not only harmed by negative administrative decisions. Digital harms start much earlier, and a negative administrative decision may only be the tip of the iceberg. In the digital age, the exclusionary effect of the power asymmetry between government and citizens may start by the choice to embrace a specific vision of a smart city with ubiquitous surveillance.

Future research should reflect upon the discrepancy between the predictive and exclusionary design and implementation of smart urban solutions and the reactive rationale of administrative law. Despite the elasticity of public law, going further, legal scholars may also want to debate the need to modify reactive administrative law paradigms and depart from the dogma of the administrative decision as the typical gateway to administrative judicial protection in many civil law jurisdictions. More importantly, administrative law should incorporate insights from different

⁶⁸ See the chapter by Margaret Warthon in this volume.

sciences to fully grasp the gap between these novel public services and the type of administrative legal protection they require. Also, this would help overcome the implicit assumptions that predictive systems have limited legal implications; the optimisation of public services is always positive and wanted by citizens; and lastly, that smart urban solutions close the gap between governments and citizens.

Putting Private Sector Responsibility in the Mix

A Business and Human Rights Approach to Artificial Intelligence

Isabel Ebert and Lisa Hsin

1 Introduction

The adverse impacts of the private sector on human rights and its responsibility for addressing such impacts have been a long-standing matter of concern in global governance.¹ With the immense corporate power and resources allocated to the development, deployment, and use of artificial intelligence (AI) technologies, this concern has gained renewed traction.² In particular, this chapter recognises concerns relating to the adverse impacts of AI on: (i) employment and labour rights;³ (ii) human rights of consumers and the world at large (such as privacy and civil and political rights);⁴ and (iii) human rights within transnational value chains. Each one of these affected groups operates transnationally as a consequence of globalised business activities.

Although neither companies nor states have yet developed a comprehensive and systematic understanding of the human rights impacts of AI, it is clear that the

¹ Florian Wettstein, *Business and Human Rights: Ethical, Legal, and Managerial Perspectives* (CUP 2022).

² See eg Office of the United Nations High Commissioner for Human Rights (OHCHR), ‘B-Tech: Bridging Governance Gaps in the Age of Technology—Key Characteristics of the State Duty to Protect’ (May 2021) <www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/b-tech-foundational-paper-state-duty-to-protect.pdf>; Shoshana Zuboff, ‘Big Other: Surveillance Capitalism and the Prospects of an Information Civilization’ (2015) 30 *Journal of Information Technology* 75; Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power* (Public Affairs 2019).

³ See Aislinn Kelly-Lyth, ‘Challenging Biased Hiring Algorithms’ (2021) 41 *Oxford Journal of Legal Studies* 899; and BBC News, ‘AI at Work: Staff “Hired and Fired by Algorithm”’ (*BBC News*, 25 March 2021) <www.bbc.co.uk/news/technology-56515827>. See also the chapter by Joe Atkinson and Philippa Collins in this volume.

⁴ UK Competitions and Markets Authority Consultation, ‘Algorithms: How They Can Reduce Competition and Harm Consumers Summary of Responses to the Consultation’ (May 2021) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/991676/Summary_of_responses_to_algorithms_paper_publish.pdf>. See also the chapter by Shu Li, Béatrice Schütte, and Lotta Majewski in this volume.

impact of AI on human rights is transnational, which requires a principled and at the same time adaptable regime to meet the challenges of new technology in a rapidly changing landscape.⁵ For instance, does the international human rights protection system provide the necessary means to deal with, for instance, racist or gender-biased AI software that supports court decisions? To what extent are companies responsible if their AI services facilitate the spread of hate speech and violence? What should corporate human rights due diligence systems look like for AI systems? These are some of the questions this chapter will seek to explore from a business and human rights (BHR) perspective.

This chapter explains that when the UN Guiding Principles on Business and Human Rights (UNGPs) were adopted in 2011, they set in motion greater recognition of human rights responsibilities for corporations in law. We set out this framework for BHR and examine key domestic legislative developments and divergences, which fit within the BHR framework. We consider how these standards apply to AI, including specific challenges. Overall, this chapter introduces the reader to how a human rights-centric approach relates to AI, and how ‘human rights due diligence’ processes, as advanced by the UNGPs, can encourage technology companies to adapt their AI operations to respect human rights.

2 Business and Human Rights: A Brief Introduction of an Emerging Framework

Since the 1970s, the United Nations (UN) has made efforts to explore the impact of private businesses on human rights. Such efforts included inquiries carried out by the Commission on Transnational Corporations from 1973 to 94; the UN Sub-Commission on Promotion and Protection of Human Rights (established in 1998) which, in turn, founded a Working Group on Transnational Corporations, but in 2003, the ‘Norms on the Responsibilities of Transnational Corporations and Other Business Enterprises with Regard to Human Rights’ (the Norms) were abandoned due to significant opposition from the business sector.⁶

A ‘fact-finding mission’ carried out by the Special Representative of the Secretary General on human rights and transnational corporations and other business enterprises, John Ruggie, took the form of a global multi-stakeholder consultation with governments, business, civil society, trade unions, and academics, which resulted

⁵ Ben Wagner, Matthias C Kettemann, and Kilian Vieth, *Research Handbook on Human Rights and Digital Technology: Global Politics, Law and International Relations* (Edward Elgar 2019); Mathias Risse and Steven Livingston, ‘The Future Impact of Artificial Intelligence on Humans and Human Rights’ (2019) 33(2) *Ethics & International Affairs* 141.

⁶ David Weissbrodt and Muria Kruger, ‘Norms on the Responsibilities of Transnational Corporations and Other Business Enterprises with Regard to Human Rights’ (2003) 97 *American Journal of International Law* 901.

in the adoption of the ‘Protect, Respect, Remedy Framework’ in 2008 by the UN Human Rights Council.⁷ The UNGPs were drafted on this basis and adopted in June 2011 through the UN Human Rights Council and have become an important cornerstone of responsible business practice today.⁸

The UNGPs adopted the ‘protect, respect and remedy framework’, which is now considered the cornerstone of BHR,⁹ and referred to as a ‘high watermark of business and human rights’.¹⁰ The UNGPs are based on three fundamental pillars:

- (i) the state’s duty to protect human rights as grounded in international human rights law;
- (ii) the corporate responsibility to respect human rights, or the ‘do no harm’ principle; and
- (iii) access to remedies for victims of corporate human rights abuses.

While other international initiatives targeting the private sector preceded the UNGPs, such as the UN Global Compact and UN Principles of Responsible Investments, the UNGPs’ direct imposition of corporate responsibility towards human rights have captured the wider attention of legal practitioners and commentators. The UNGPs provide a soft law framework that lays out the expectations towards business to respect human rights and spell out the state’s role in requiring business to do so.

To elaborate, the first pillar requires states to meet their duty to protect by taking ‘appropriate steps to prevent, investigate, punish and redress human rights abuse through effective policies, legislation, regulations and adjudication’ (UNG P 1), and that states ‘should consider a smart mix of measures—national and international, mandatory and voluntary—to foster business respect for human rights’ (UNG P 3).

The second pillar requires businesses to respect human rights and calls on companies to avoid infringing on the human rights of others and address adverse human rights impacts they are involved with (UNG P 11). The UNGPs introduce the concept of ‘human rights due diligence’ to prevent and mitigate risks to people, not risks to business, even though these may converge. To meet their responsibility to respect human rights, companies are asked to put in place ‘policies and processes appropriate to their size and circumstances’, including a ‘human rights due

⁷ Florian Wettstein, ‘Normativity, Ethics, and the UN Guiding Principles on Business and Human Rights: A Critical Assessment’ (2015) 14 *Journal of Human Rights* 162.

⁸ Human Rights Council/UNGA, ‘Protect, Respect and Remedy: A Framework for Business and Human Rights’ (7 April 2008) UN Doc A/HRC/8/5; OHCHR, ‘Guiding Principles on Business and Human Rights: Implementing the United Nations’ “Protect, Respect and Remedy” Framework’ (21 March 2011) UN Doc A/HRC/17/31.

⁹ Florian Wettstein, ‘From Side Show to Main Act: Can Business and Human Rights Save Corporate Responsibility’ in Dorothee Baumann-Pauly and Justine Nolan (eds), *Business and Human Rights: From Principles to Practice* (Routledge 2016) 78.

¹⁰ Alex Newton, *The Business of Human Rights: Best Practice and the UN Guiding Principles* (Routledge 2019).

diligence process to identify, prevent, mitigate and account for how they address their impacts on human rights' (UNGPs 15). Effective due diligence is context-dependent and flexible, but its aim is to embed human rights into corporate policies and management systems and enable companies to remediate adverse impacts that they cause or contribute to.

The third pillar constitutes the right to an effective remedy for violations of human rights, as enshrined in international human rights law and confirmed in a wide range of international law instruments, including key components of the International Bill of Rights.¹¹ The legal obligations of states to provide access to effective remedies for business-related human rights harms, including human rights harms associated with the development and use of digital technologies by companies, form part of the state duty to protect human rights, as reiterated by the UNGPs.¹²

Vehicles for seeking and delivering remedies for business-related human rights harms required by the UNGPs can be categorised into the following main types: (i) judicial mechanisms, such as domestic courts, regional courts, regional and international human rights bodies; (ii) state-based non-judicial mechanisms, such as regulators, ombudspersons, inspectorates, public complaints handling bodies, National Contact Points under the OECD Guidelines; and (iii) remediation mechanisms developed and administered by private entities, including industry associations or multi-stakeholder groups.

Company-based grievance mechanisms can serve as an important vehicle through which technology companies can directly address and remedy human rights harms. The UNGPs ask business enterprises to 'establish or participate in effective operational-level grievance mechanisms' for affected individuals and communities to enable early and direct resolution of grievances arising from adverse human rights impacts.

To implement effective remedies through the three types of mechanisms set out above, the UNGPs have to be translated into the domestic legal context, with the aid of local law. The idea is that by incorporating the UNGPs into domestic law, a 'patchwork' of legal standards could form over transnational businesses activities while referring to a consistent principled-based approach grounded in the UNGPs. Corporate irresponsibility and impunity which have long benefitted from the governance gap between public international law (applicable only to states), and domestic regimes (limited to territory), should, in theory, no longer be able to escape scrutiny.

¹¹ See, in particular, the Universal Declaration of Human Rights (UDHR), art 8; and the International Covenant on Civil and Political Rights (ICCPR), art 2. See further UNGA Res 60/147 (16 December 2005) UN Doc A/RES/60/147.

¹² OHCHR (n 2).

3 AI and Business: Domestic and Regional Human Rights Initiatives

Having established that domestic and regional initiatives are necessary for implementation of the UNGPs, the next question is should the UNGPs be incorporated? The UNGPs envisage wide application to capture the transnational use of AI in global value chains, such as during the development and deployment of AI products and services. Some states have adopted legislation incorporating ‘human rights due diligence’ (HRDD). Others opted for limited disclosure obligations for a smaller subset of human rights violations, such as the Modern Slavery Act 2015 in the United Kingdom (UK). In this section, we discuss how legislation framed as BHR can give effect to the UNGPs, while offering a non-exhaustive account of recent state-based legislative developments with extraterritorial effects. In doing so, we illustrate the variety of mechanisms available and divergences in legislative and policy approaches. In the end, we suggest that HRDD may not be a panacea, but it could be the most effective mechanism for the challenges imposed by AI to date. We argue that the process-oriented character of a human rights due diligence approach, based on a thorough assessment of potential adverse impacts on human rights stemming from AI, can capture the fast-evolving nature of technological innovation, account for the individual usage of AI in various organisational contexts and at the same time provide the appropriate approach to be responsive in parallel to being principled in requiring high-quality standards of responsible business conduct.

We begin chronologically, with the UK’s Modern Slavery Act 2015 (MSA), which at the time, was heralded as a breakthrough in corporate accountability. Following its enactment, the law generated considerable public awareness, shifted corporate priorities, and created momentum for legislation elsewhere, including Australia, which adopted its own Modern Slavery Act—largely modelled on the MSA—in 2018.¹³ Both statutes rely on the effects of ‘trickle down’ to impose obligations extraterritorially, but narrowly focus on ‘modern slavery’ instead of human rights more broadly. The definition of ‘modern slavery’ encompasses long-standing criminal offences such as human trafficking, forced labour, and slavery.¹⁴ Canada and New Zealand have also been working on proposals for supply chain transparency, but similarly limited to modern slavery.¹⁵

In contrast, the United States (US) has adopted an industry-specific approach to its disclosure obligations. In addition to the California Transparency in Supply Chains Act 2010, the US has existing laws designed to target corporate sourcing

¹³ Modern Slavery Act 2018 (Australia).

¹⁴ Modern Slavery Act 2015, Pt 1 (UK).

¹⁵ Bill S-211 (Canada) <www.parl.ca/DocumentViewer/en/44-1/bill/S-211/first-reading>; New Zealand Ministry of Business, Innovation and Employment, ‘Consultation on Modern Slavery and Worker Exploitation’ <www.mbie.govt.nz/have-your-say/modern-slavery/>.

practices. Section 1502 of the Dodd–Frank Act of 2010 requires publicly traded companies in the US with products containing ‘conflict minerals’ such as gold, tin, tungsten, and tantalum (some of which are commonly used in the production of electronic goods), to report on the origins of those materials.¹⁶ The purpose of the law is to expose the underlying source of trade in ‘conflict minerals’.

The French *Loi de Vigilance* is the first legislative example of a general mandatory due diligence requirement for human rights and environmental impacts. The law imposes a ‘duty of vigilance’ on certain large French companies (employing 5,000 employees in France or 10,000 globally), and the law extends to activities of French companies’ subsidiaries ‘with which the company maintains an established commercial relationship’. Following this test, companies are held to have a legal duty, which can give rise to legal action by civil society.¹⁷ The French law consists less of a reporting requirement, as its primary goal is to require companies to implement a ‘vigilance plan’ which is supposed to include reasonable measures to adequately identify serious violations of human rights and impacts on the environment.

The French law is a significant milestone and departure from the existing approach. The result is that companies should look at their supply chains and find out what is really happening. In contrast, the UK provision is a reporting requirement, which encourages companies to investigate, but does not require companies to undertake any specific action. The question now on the international stage is one of legislative design: are due diligence requirements, as advanced by the UNGPs, the way forward, or do we rather need transparency requirements?

Overall, transparency and light-touch disclosure regimes such as the MSA have been criticised for insufficiently holding corporations to account and for being too limited in scope.¹⁸ Consequently, lobbying for mandatory HRDD on all human rights and environmental issues has gained momentum. In late 2021, civil society and some large businesses called on the UK Government to adopt a Business, Human Rights and Environment Act, which would require businesses to conduct mandatory HRDD.¹⁹ However, the level of enforcement for HRDD regimes appears to have fallen short of expectations as well. In December 2019, the French non-governmental organisation (NGO) Friends of the Earth sued French oil giant Total over the impact of its commercial activities in Uganda under the

¹⁶ Dodd–Frank Act of 2010, s 1502 (US).

¹⁷ *Friends of the Earth et al v Total Energy*, Nanterre High Court (2021) <http://climatecasechart.com/climate-change-litigation/wp-content/uploads/sites/16/non-us-case-documents/2021/2021216_NA_decision-1.pdf>.

¹⁸ Independent Review of the Modern Slavery Act 2015: Final Report (CP100) (2019) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/803406/Independent_review_of_the_Modern_Slavery_Act_-_final_report.pdf> 40.

¹⁹ Business and Human Rights Resource Centre, ‘UK: Businesses and Investors Call for New Human Rights Due Diligence Law’ (21 October 2021) <www.business-humanrights.org/en/latest-news/uk-businesses-and-investors-call-for-new-human-rights-due-diligence-law/#:~:text=In%20October%202021%2C%2036%20companies,an%20the%20British%20Retail%20Consortium>.

Loi de Vigilance.²⁰ However, the Nanterre High Court accepted the arguments of Total's lawyers and decided that it did not have jurisdiction to decide the matter, deferring to the Commercial Court instead. As these issues remain unresolved, commentators are starting to suggest that the French law is also ineffective. Some practitioners expressed cynicism that the law is nothing more than a reporting requirement, which renders it to be of the same level of effectiveness as the MSA.²¹ While these regimes do not mention AI impacts on human rights specifically, the pervasive nature of AI in global value chains means that the use of AI should and would inevitably be subject to greater scrutiny.

Nevertheless, the trend towards mandatory human rights due diligence legislation continues, including in the field of technology and AI. Domestic and regional legislation, such as the UK's proposed Online Harms Bill,²² and the European Union's (EU) Digital Services Act (DSA),²³ have emerged to address adverse human rights effects of AI.

In 2021, both Germany and Norway adopted new laws mandating HRDD by companies. In Germany, the Supply Chain Due Diligence Act entered into force on 1 January 2023.²⁴ This affirms Germany's intention to comply with EU due diligence obligations. The Norwegian Transparency Act applies to large enterprises (based on sales revenue or if they have more than fifty full-time employees) resident in Norway which offer goods and services in or outside of Norway.²⁵ The Norwegian Act includes fundamental human rights—referencing the OECD Guidelines for Multinational Enterprises—and decent working conditions, whereas the German law also includes environmental risks.

At the EU level, after years of debate between EU parliamentarians, civil society, companies, and EU member states, the European Commission issued its Proposal for a Directive on Corporate Sustainability Due Diligence to tackle human rights

²⁰ Friends of the Earth International, 'Total Abuses in Uganda: French High Court of Justice Declares Itself Incompetent in Favour of the Commercial Court' (30 January 2020) <www.foei.org/no-category/total-abuses-uganda-french-high-court-of-justice-declares-itself-incompetent-duty-vigilance-law>.

²¹ Business and Human Rights Resource Centre, 'Three Years On: How Has the French Law on the Corporate Duty of Vigilance influenced Human Rights Due Diligence?' (14 May 2020) <www.business-humanrights.org/en/three-years-on-how-has-the-french-law-on-the-corporate-duty-of-vigilance-infused-human-rights-due-diligence>.

²² In the UK, the Online Safety Bill is a significant piece of draft legislation that, when in force, will make provision for the regulation of certain internet services, with the primary aim of tackling illegal and harmful content disseminated online.

²³ The European Commission has proposed two legislative initiatives to upgrade rules governing digital services in the EU: the Digital Services Act and the Digital Markets Act. The Commission made the proposals in December 2020 and political agreements were reached on the following dates: 25 March 2022 (Digital Markets Act) and 23 April 2022 (Digital Services Act).

²⁴ Act on Corporate Due Diligence in Supply Chains 2021 (Germany) <<https://perma.cc/8JUX-ET2Q>>. See <www.loc.gov/item/global-legal-monitor/2021-08-17/germany-new-law-obligates-companies-to-establish-due-diligence-procedures-in-global-supply-chains-to-safeguard-human-rights-and-the-environment>.

²⁵ Act Relating to Enterprises' Transparency and Work on Fundamental Human Rights and Decent Working Conditions (Transparency Act) 2021 (Norway) <<https://lovdata.no/dokument/NLE/lov/2021-06-18-99>>.

and environmental impacts across global value chains.²⁶ The Directive would impose a corporate due diligence duty on large EU and third-country companies, and smaller companies in certain ‘high-risk’ sectors, to identify and take steps to remedy actual and prevent or mitigate potential adverse impacts on human rights in the companies’ own operations, and their subsidiaries and value chains.

Having summarised and identified the prominent domestic BHR developments in law, let us now reflect upon the implications of these developments on AI specifically. NGOs and civil society stakeholders have called for the UNGPs to be incorporated into domestic law for companies developing and deploying AI. However, exactly how to marry AI and HRDD is not a clear-cut exercise of legislative design.

Given the vast scope of possible AI applications, it can be challenging to define obligations *ex ante*. At the same time, an advantage of the UNGP approach is precisely its principled but adaptable approach: the UNGPs and HRDD processes can be adapted and tailored to specific companies and business activities. After all, the human rights impacts connected to the business are assessed and mitigated on an ongoing basis and hence such an approach appears fruitful along the AI life cycle, from development to deployment to end-use. Perhaps more than any other industry or business activity, AI systems, products, and services can cause or contribute to widespread and unforeseen adverse impacts. This means that monitoring, engagement, and the willingness to take human rights seriously must be core purposes of companies developing and deploying AI.²⁷

To that end, and from a BHR standpoint, a process-oriented legislative design focused on impacts on rightsholders would be favourable to tackle the specific human rights challenges connected with AI. Ongoing discourse at the international level also proposes the following key elements that a meaningful AI regulation with an emphasis on rights-respecting company conduct should entail:²⁸

- (i) a broad view on human rights obliging companies to identify, assess and mitigate a wide range of human rights impacts connected to their AI technologies and services, while requiring a prioritisation of action by businesses based on the saliency/severity of the potential impact(s) on the rightsholders;

²⁶ European Commission, ‘Proposal for a Directive of the European Parliament and of the Council on Corporate Sustainability Due Diligence and amending Directive (EU) 2019/1937 COM(2022) 71 final, 2022/0051(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0071>>.

²⁷ Colin Mayer, ‘The Future of the Corporation: Towards Humane Business’ (2018) 6 *Journal of the British Academy* 1.

²⁸ Isabel Ebert and Ana Beduschi, ‘Regulating Business Conduct in the Technology Sector: Gaps and Ways Forward in Applying the UNGPs’ (2022) <https://iwe.unisg.ch/fileadmin/user_upload/HSG_ROOT/Institut_IWE/Forschung/Publikationen/Papers/Research_Brief_-_Regulating_tech_business_conduct__UNGP_April_2022.pdf>.

- (ii) a consistent application of HRDD terminology across legislative design to avoid diluting the expectations towards company conduct as set out by the UNGPs;
- (iii) a value chain focus across the full business sphere, to ensure capturing the impacts of AI along the full life cycle of a product or service;
- (iv) accompanying measures and proper enforcement in order to ensure robust implementation;
- (v) a process-oriented character of the legislation relating to the expectations for businesses to meet, which allows addressing impacts on rightsholders as they emerge;
- (vi) building on genuine stakeholder engagement to understand the needs of affected stakeholders (users and non-users of AI); and
- (vii) clear provisions for access to remedies, which also implies accountable, traceable, and retractable development and usage of AI.

4 Applying HRDD to Artificial Intelligence

Given the momentum behind BHR initiatives and the ongoing ‘tech lash’,²⁹ it is timely to further explore the intersection of AI companies and the UNGPs.³⁰ BHR initiatives involving multiple stakeholders have added pressure on technology companies to enhance transparency and to adapt their practices to align with BHR principles. In particular, in 2019, the Office of the High United Nations Commissioner for Human Rights (OHCHR) launched a flagship project called the ‘B-Tech Project’, which seeks to provide authoritative guidance and resources for implementing the UNGPs in the technology area. The project was launched after consultations with civil society, business, states, and experts, about its scope and it provides guidance for companies and key stakeholders on a regular basis to promote HRDD as an overarching framework.

As the UNGPs apply broadly to all business enterprises, all enterprises within the technology industry including those developing, deploying, and using AI regardless of ‘size, sector, operational context, ownership and structure’ (UNGP 14) are subject to it. The UNGPs also apply to any enterprise from any other sector that makes use of an AI product, service, and/or solution. Such enterprises must, therefore, conduct HRDD to ensure that the procurement and use of AI technologies does not result in adverse human rights impacts. When states are the end-users

²⁹ Ronald J Deibert, ‘The Road to Digital Unfreedom: Three Painful Truths about Social Media’ (2019) 30(1) *Journal of Democracy* 25; Caroline Criado Perez, *Invisible Women: Exposing Data Bias in a World Designed for Men* (Vintage 2019); Cathy O’Neil, *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy* (Broadway Books 2017).

³⁰ R McCorquodale and M Tse, ‘Artificial Intelligence Impacts: A Business and Human Rights Approach’ (2021) 26 *Communications Law* 11.

of AI, they must act in ways that are consistent with their human rights obligations. The UNGPs recognise—and are flexible enough to accommodate companies involved with AI, which are expected to tailor their HRDD process to the complexity that comes with the size of the business enterprise—the risk of severe human rights impacts, and the nature and context of their operations. The context in which AI is applied in products, services, and solutions of the technology industry varies dramatically, spanning, for example, social media platforms, search engines, geolocation tools, as well as facial recognition systems.

HRDD should take place early and often throughout processes of product design, development, and use of AI systems, products, and services.³¹ The HRDD process can be largely broken down into four steps: (i) identifying and assessing impacts to gauge the nature and extent of human rights risks; (ii) acting to prevent and mitigate risks to people, including via integration within internal functions and processes; (iii) tracking of effectiveness of risk mitigation responses over time; and (iv) appropriate communication of performance with respect to addressing human rights impacts.³² This approach to HRDD builds on existing corporate commitments to respect human rights, which are often embedded through approval at board/executive level.

Additionally, the UNGPs call on companies to set up or participate in mechanisms to hear and address grievances about human rights impacts stemming from or being linked to their business activities. For example, when it comes to identifying and assessing human rights impacts connected with AI, it might be that a company risks using datasets that are not representative, which can result in bias towards, for instance, underrepresented minorities. Such bias could lead to subsequent flaws in automated detection. In the case of facial recognition technology, there have been concerns that it does not work accurately for people of colour,³³ which results in false positives, *inter alia*, in policing.³⁴ To prevent adverse human rights impacts, companies should conduct data testing before large-scale product roll-outs and convene stakeholder focus groups, including with vulnerable groups. As HRDD is an ongoing process, efforts should be made to track effectiveness. Companies need to undertake regular review processes of AI systems, to minimise the likelihood of discrimination against, in particular, minorities, women, and people of colour. Companies should then report publicly about the steps they have taken to address human rights risks—not just reputational risks. As shown previously in regard to the MSA, this could be done by way of ‘transparency reporting’, which captures, on

³¹ OHCHR, ‘B-Tech Project: Key Characteristics of Business Respect for Human Rights—A B-Tech Foundational Paper’ (2020) <www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/key-characteristics-business-respect.pdf>.

³² *ibid.*

³³ J Buolamwini, ‘When the Robot Doesn’t See Dark Skin’ *The New York Times* (21 June 2018) <www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html>.

³⁴ See the chapter by Valentina Golenova in this volume.

an annual basis, human rights risks identified, and preventative, redress measures implemented by companies.

HRDD, in our view, is suited to AI for various reasons, including: (i) the process-oriented character allowing for evolving technologic innovation; (ii) the responsiveness to contextual factors that shape AI impacts; and (iii) principled-based expectations providing for enhanced clarity among stakeholders. In the following sections, we discuss how AI challenges human rights, and how HRDD legislation may provide part of an answer to these challenges.

While the term 'AI' is often used as a hypernym, with the exact technological configurations depending on sectoral focus and the ecosystem of the AI usage, specific implications on BHR can be understood in terms of a 'chronology'.³⁵ The AI life cycle that can be broadly divided into early development, testing, and roll out.³⁶ Let us note the key challenges to human rights at each stage of the AI process.

At the development stage, AI relies heavily on large data sets and data training; therefore, it is crucial to follow the data life cycle in order to detect unintended consequences and blind spots with regard to potential adverse impacts on human rights. Consultation with the product team and engineers developing AI, as well as policy teams, is essential to bridge organisational silos before product roll-outs. Similarly, engagement with affected stakeholder groups prior to market entry and beyond can reveal insightful information to adjust products in a rights-respecting manner. More generally, transparent communication to users about what data is collected and how is essential.

Businesses models and corporate practices of companies using and deploying AI, which focus on profit maximisation, can contribute to systematic, negative impacts on a range of human rights—for instance, by optimising for the monetisation of behavioural data and, hence, maximisation of user data.³⁷ To minimise the likelihood of such impacts, key performance indicators could be reimaged to reward activities that prevent or mitigate human rights harms, while cross-departmental training can cut across highly segregated business models, so as to encourage more communication between employees and suppliers.

When testing the product in new markets, AI systems, products, and services must be subject to scrutiny. It is essential to stress-test and where necessary improve the design of AI technologies in ways that demonstrably minimise the risks of severe human rights harms, as opposed to solely optimising technologies for revenue maximisation. For example, datasets feeding into AI analysis should be

³⁵ Global Network Initiative and BSR, 'Understanding Human Rights Due Diligence (HRDD) Under an Ecosystem Lens' (2022) <<https://eco.globalnetworkinitiative.org>>.

³⁶ For a comprehensive account, see the chapter by Martina Šmuclerová, Luboš Král, and Jan Drchal in this volume.

³⁷ United Nations, Office of the High Commissioner for Human Rights, B-Tech Project (2020): Addressing Business Model Related Human Rights Risks <https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech_Foundational_Paper.pdf>.

representative in proportion to the local population context, while accounting for the protection of vulnerable groups. This also entails scrutinising plans for testing and expansion in new markets, with a focus on whether the local context exacerbates business model human rights risks (eg a conflict might be fuelled by certain AI-based products recommending polarising content to make users stay longer on platforms and hence increase the potential for the monetisation of user data). Furthermore, collective action engagement with peers, professional associations, customers, civil society, and government can help to develop and implement more robust rights respecting standards of business conduct and technological design. Overall, in line with the immense market power, companies have a responsibility to ensure they play a constructive role in processes to develop laws and regulations aimed at increasing human rights protections.

At the levels of deployment and roll-out, further challenges include impact on third parties who are not the users of technologies, such as through the AI-facilitated spread of hate speech and incitement of violence on social media platforms. In practice, AI impact is subject to the scope of its application. Thus, the impact of the final roll-out can have rather different results to the earlier tests conducted over smaller samples. Furthermore, certain rightsholder groups can be systematically discriminated at the roll-out stage by AI if this was not detected adequately in testing.

Hypothetically speaking, if a company selling or licensing AI solutions to a government border agency becomes concerned with the behaviour of the government agency, according to the UNGPs that company would need to take reasonable steps to seek to prevent or mitigate such impact. The HRDD process is essential for identifying these issues. The UNGPs would justify the company's decision to restrict the use of certain features or engage with the agency about the company's concerns. Otherwise, the AI company could be seen to be facilitating the continuance of the agency's bad behaviour and contributing to human rights harm.³⁸ This in turn can give rise to potential legal claims, including civil action in some jurisdictions.³⁹

Across the stages of the AI life cycle, the human rights track record of 'Big Tech', which are often at the forefront of developing and deploying AI products and services due to their expansive market power and well-resourced AI innovation hubs, remains questionable. A recent report of the Human Rights Council highlighted ongoing challenges and gaps in the implementation of the corporate responsibility to respect human rights in the technology sector, being the biggest adopters of AI.⁴⁰ One example included the current lack of practices regarding stakeholder

³⁸ OHCHR (n 31).

³⁹ See Ekaterina Aristova and Ugljesa Grusic, *Civil Remedies and Human Rights in Flux: Key Legal Developments in Selected Jurisdictions* (Bloomsbury 2022).

⁴⁰ OHCHR, 'The Practical Application of the Guiding Principles on Business and Human Rights to the Activities of Technology Companies' (21 April 2022) UN Doc A/HRC/50/56.

engagement as part of HRDD. Another issue described was insufficient access to affected stakeholders, in particular in the Southern hemisphere, as the majority of large technology companies do not necessarily entertain direct policy relationship with vulnerable groups. An additional concern was related to the mixed level of company performance with regard to transparency reporting and communication. Where companies do engage, the style of their engagement has been found too excessive, causing engagement fatigue and resource strain on certain groups, through practices such as onerous demands and repeated requests for input from a small number of civil society organisations. These concerns have led some experts to call for standardisation of transparency reporting.⁴¹ Also, publicly accessible rankings, such as Ranking Digital Rights, can consolidate information on the performance of technology companies, including corporate activities in relation to human rights.

5 Conclusion

AI technology transcends boundaries. As Phil Bloomer of the Business and Human Rights Resource Centre pointed out, ‘artificial intelligence, automation, and the gig economy can free us from drudgery, enrich our leisure, and build societies of shared prosperity. [But] [t]hey have equal potential to create mass unemployment, hollow-out lives, and worsen inequality. Putting human rights at the core of this new wave of technology in global markets will help define which road we choose.’⁴² The question is *how*?

In this chapter, we highlighted the pervasive nature of AI in value chains, the key role that businesses play in deploying and adopting the use of AI. We have illustrated that while many businesses operate transnationally, laws are traditionally bound by geography and jurisdiction. The significant contribution of the UNGPs and HRDD is to impose an obligation on corporations to respect human rights. The adaptability of this principle-based, process-oriented approach leverages the business models and existing resources, to apply even to the most groundbreaking technology.

However, the UNGPs must be adopted into domestic law to take effect. As such, initiatives and regimes discussed in this chapter are only the nascent beginning. It appears that at least in Europe, the current legislative design trend for the digital economy goes towards process-based regulation.⁴³ Regulations, which are mindful of different corporate risk profiles, should be carried out with regard to the UNGPs and complement existing corporate BHR processes. Regulatory initiatives

⁴¹ Christopher Parsons, ‘The (in)effectiveness of Voluntarily Produced Transparency Reports’ (2019) 58 *Business & Society* 103.

⁴² See <www.business-humanrights.org/de/schwerpunkt-themen/technology-human-rights/>.

⁴³ Giovanni De Gregorio and Pietro Dunn, ‘The European Risk-Based Approaches: Connecting Constitutional Dots in the Digital Age’ (2022) 59(2) *Common Market Law Review* 473.

will need to assess the responsibilities of companies that develop or deploy AI, such as when it comes to data protection in AI processes (eg the EU General Data Protection Regulation),⁴⁴ the role of AI in content governance (eg the EU Digital Services Act)⁴⁵ but also the characteristics and set-up of AI systems more in general (eg the EU AI Act).⁴⁶

Given the current trajectory, companies developing and deploying AI will increasingly find themselves subject to overarching international and regional obligations and national, domestic BHR due diligence legislation. While this will increase standards and encourage corporate adherence to human rights, it also highlights the importance of alignment between domestic and international standards regarding business conduct relating to AI products and services. Bearing in mind that ineffective application of the UNGPs can dilute their significance,⁴⁷ moving forward, regulators must seriously consider the use of AI transnational business activities and potential human rights impacts. Only through such an approach will it be possible to prevent harm and systematically integrate human rights into transnational AI value chains.

⁴⁴ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (General Data Protection Regulation/ GDPR).

⁴⁵ European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC' (DSA) COM(2020) 825 final, 2020/0361(COD), arts 26–27.

⁴⁶ European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts' (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

⁴⁷ Ebert and Beduschi (n 28).

Artificial Intelligence Human Rights Impact Assessment

Alessandro Ortalda and Paul De Hert*

1 Introduction

In the last decades, technology has evolved dramatically. Not only has computing power consistently risen (following the 1965 prediction by Gordon Moore),¹ but the uptake of digital means has also exploded, thus bringing technology to every possible avenue of human activity.² Among technologies, artificial intelligence (AI) is becoming one of the most relevant, growing in terms of capabilities³ and adoption.⁴ As AI becomes more embedded in society, it is important to assess the consequences of its use on people and their human rights. Recent examples from the Netherlands (the *SyRI* scandal⁵ and the *Toeslagenaffaire*)⁶ demonstrate that the use of AI has impacts that can be invasive to human rights. These impacts vary according to the specific sector or application of AI,⁷ but also according to the category of users. For instance, the United Nations Human Rights Council recently

* The authors would like to thank Alessandra Calvi for the precious support and feedback received during the drafting of this contribution.

¹ Gordon Earl Moore, ‘Cramming More Components onto Integrated Circuits’ (1965) 38 Electronics 114.

² This is continuing to this day (possibly accelerating) thanks to multiplication of devices and generation of new data. See European Commission, ‘Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions. A European Strategy for Data’ COM(2020) 66.

³ For instance, thanks to application of AI in the medical sector, according to the technology research firm Gartner, ‘by 2025, more than 30% of new drugs and materials will be systematically discovered using generative AI techniques’. See Meghan Rimol, ‘Gartner Identifies Key Emerging Technologies Spurring Innovation Through Trust, Growth and Change’ (*Gartner*, 23 August 2021).

⁴ This trend has also consolidated around recent events. Artificial intelligence experienced a peak growth in the wake of the COVID-19 crisis. See Joe McKendrick, ‘AI Adoption Skyrocketed Over the Last 18 Months’ (*Harvard Business Review*, 27 September 2021) <<https://hbr.org/2021/09/ai-adoption-skyrocketed-over-the-last-18-months>>.

⁵ See Adamantia Rachovitsa and Niclas Johann, ‘The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case’ (2022) 22 Human Rights Law Review 1.

⁶ See Melissa Heikkilä, ‘AI: Decoded: A Dutch Algorithm Scandal Serves a Warning to Europe—The AI Act Won’t Save Us’ (*Politico*, 30 March 2022) <www.politico.eu/newsletter/ai-decoded/a-dutch-algorithm-scandal-serves-a-warning-to-europe-the-ai-act-wont-save-us-2/>.

⁷ Josephine Yam and Joshua August Skorburg, ‘From Human Resources to Human Rights: Impact Assessments for Hiring Algorithms’ (2021) 23 Ethics and Information Technology 611.

brought attention to the possible negative effects for persons with disabilities and their enjoyment of human rights.⁸

Not only technology changes. Human rights are equally undergoing a period of change. Already in the late seventies, the European Court of Human Rights (ECtHR) acknowledged in the *Tyrrer v the United Kingdom* judgment that the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) is ‘a living instrument which … must be interpreted in the light of present-day conditions’.⁹ This position has been upheld also in subsequent judgments.¹⁰ For some, relying on judicial interpretation only is not sufficient and they advocate for broadening the human rights framework with new rights.¹¹ There are indeed situations and interests that are not covered or protected by the current list of recognised rights.¹² Policy actors like the European Commission are therefore at work to translate rights into the current digital landscape and to make them fit for it.¹³

The growing relevance of AI and the permanent evolution of human rights frameworks make assessing AI impacts on human rights more complex and more pressing. The goal of the present contribution is to understand how the practice of human rights impact assessment in the context of AI systems can be successfully implemented by concerned stakeholders.¹⁴ To do so, this chapter tackles first the meaning of human rights impact assessment, and how this practice is still relatively rare in the legal framework. Impact assessment is looked at through the lenses of data protection law, with particular regard to the Data Protection Impact Assessment (DPIA) found in the European Union (EU) General Data Protection Regulation (GDPR).¹⁵ The GDPR is one of the most advanced legal documents on data protection and despite being a relatively young one, it has already reached a relevance that goes beyond European borders. We contend that impact assessment

⁸ United Nations Human Rights Council, ‘Rights of Persons with Disabilities: Report of the Special Rapporteur on the Rights of Persons with Disabilities’ UN Doc A/HRC/49/52 (2021). See also the chapter by Antonella Zarra, Silvia Favalli, and Matilde Ceron in this volume.

⁹ *Tyrrer v the United Kingdom* App no 5856/72 (ECtHR, 25 April 1978), para 31.

¹⁰ For a historical account, see George Letsas, ‘The ECHR as a Living Instrument: Its Meaning and Its Legitimacy’ <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=%202021836>.

¹¹ See eg Andreas von Arnauld, Kerstin von der Decken, and Mart Susi (eds), *Cambridge Handbook of New Human Rights: Recognition, Novelty, Rhetoric* (CUP 2020).

¹² See eg Paul Quinn, ‘Crisis Communication in Public Health Emergencies: The Limits of “Legal Control” and the Risks for Harmful Outcomes in a Digital Age’ (2018) 14 Life Sciences, Society and Policy 4.

¹³ European Commission, ‘European Declaration on Digital Rights and Principles for the Digital Decade’ COM(2022) 28 final.

¹⁴ Indeed, scholars investigated how human rights impact assessment in the context of AI sometimes fail to meaningfully fulfil their scope when put into practice. See eg Mark Latonero and Aaina Agarwal, ‘Human Rights Impact Assessments for AI: Learning from Facebook’s Failure in Myanmar’ [2021] Carr Center for Human Rights Policy Harvard Kennedy School, Harvard University <<https://carrcenter.hks.harvard.edu/files/cchr/files/210318-facebook-failure-in-myanmar.pdf>>.

¹⁵ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L119/1 (GDPR).

within the meaning of the GDPR has reached a maturity that can inform the practice of assessing human rights impacts in the context of AI systems and represents a valuable use case outside of Europe as well. However, DPIA is mandated only when a personal data processing occurs. Therefore, even if an AI system could potentially introduce impacts to human rights, there is no obligation to assess such impact, absent a personal data processing. The analysis of the recent proposal for an EU Artificial Intelligence Act (AI Act)¹⁶ from the European Commission confirms this point, despite the effort currently ongoing at the European Parliament to cover this gap¹⁷.

This normative setup fails to capture the complexities of the current technological scenario. For instance, Schermer raises this point when talking about the advent of ambient intelligence. Accordingly, data will be recorded in multiple places and will often only be indirectly related to a person. For example, Radio-Frequency Identification (RFID) tags can be completely anonymised, but if a surveillance camera linked to an RFID reader shows who is carrying the tag, this can still be a case of personal data. Such indirect links and compilations of various types of (anonymous) data will become increasingly common, potentially making the concept of personal data (protection) no longer correspond to the technical and social reality.¹⁸ Thus, the present contribution investigates how the DPIA approach sanctioned by the GDPR can be extended beyond the boundaries of the data protection to include situations that do not involve personal data processing. The result—which we named Artificial Intelligence Human Rights Impact Assessment (HRIA-AI)¹⁹—is a voluntary approach to assess the human rights impacts of AI. Concerned stakeholders can employ this also when personal data processing occurs and it is flexible enough to respond to the challenges of the changing technology and evolving human rights landscape.

In section 2, we contend that, despite human rights impact assessment still being a relatively underdeveloped aspect in the legal framework, it has reached a

¹⁶ European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts’ (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

¹⁷ In a compromise text from mid-May 2023, the EU Parliament voted in favour of introducing fundamental rights impact assessments as a pre-requisite for high-risk AI systems. European Parliament, ‘DRAFT Compromise Amendments on the Draft Report. Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts’ <<https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf>>.

¹⁸ Bart Schermer, ‘Ambient Intelligence, Persoonsgegevens En Consumentenbescherming’ (ECP Platform voor de InformatieSamenleving 2008) <<https://zoek.officielebekendmakingen.nl/kst-31200-XIII-57-b1.pdf>> 32.

¹⁹ Literature often uses the term ‘Algorithmic Impact Assessment’ (AIA) to indicate impact assessment in the context of AI. We adopt the more generic term of HRIA-AI since AIA suggests a narrower view, more concerned with how the underlying algorithm is designed. In our opinion, this does not capture effectively all the aspects concerning how an AI system is used.

certain degree of maturity in the context of data protection law. Given this maturity, we maintain that impact assessment taken from data protection law (DPIA) can somehow be already considered a HRIA-AI. However, being applicable only in the context of personal data processing, it falls short of addressing human rights impacts when personal data processing does not occur. Thus, we recognise the value of DPIA as a model to design a HRIA-AI that can be adopted also outside of personal data processing. In section 3 we look at Data Protection Impact Assessment, to understand its main characteristics. In section 4, we put forward our proposal for a HRIA-AI, strictly following the structure of DPIA described in section 4. Finally, in section 5, we wrap up and summarise our contribution and identify the need for further work on how foresight techniques can be used in the context of human rights impact assessment.

2 The Role of DPIA in Assessing Human Rights Impacts of Artificial Intelligence

Since the 1948 Universal Declaration on Human Rights (UDHR), legal and voluntary human rights instruments have been produced across different regions,²⁰ or concerning specific groups of people.²¹ These instruments set out the obligations that concerned stakeholders must comply with to ensure the respect of human rights. Even if it has been said that impact assessment ‘should not be seen as a simplistic compliance exercise’,²² it represents a way to operationalise the rules enshrined in these norms. For instance, Principle 17 of the 2011 United Nations Guiding Principles on Business and Human Rights (UNGPs) states that ‘to identify, prevent, mitigate and account for how they address their adverse human rights impacts, business enterprises should carry out human rights due diligence [including the assessment of] actual and potential human rights impacts’.²³ Though obligations to perform impact assessment remains scarce, they have started to emerge in recent years, primarily in privacy and data protection law. According to scholars, their origins can be traced back to the older practices of technology impact assessment and environmental impact assessment,²⁴ and mentions of impact

²⁰ See eg Charter of Fundamental Rights of the European Union [2000] OJ C364/01; and the African Charter on Human and Peoples’ Rights 1981 (CAB/LEG/67/3 rev 5, 21 ILM 58 [1982]) 18.

²¹ See eg the Convention on the Rights of the Child (adopted 20 November 1989) UNGA Res 44/25 (CRC).

²² Office of the United Nations High Commissioner for Human Rights (OHCHR), ‘Key Characteristics of Business Respect for Human Rights’ (2020) <<https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/key-characteristics-business-respect.pdf>> 1.

²³ United Nations Human Rights Council, ‘Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework’ UN Doc A/HRC/17/31 (2011).

²⁴ See eg Roger Clarke, ‘Privacy Impact Assessment: Its Origins and Developments’ (2009) 25 Computer Law & Security Review 123.

assessments in the context of data protection go back at least to the 1970s, although structured methodologies to perform them began to emerge consistently only in the second half of the 1990s.²⁵

DPIA²⁶ is mandated by international treaties, such as article 10(2) of Convention 108+,²⁷ as well as EU legal instruments, such as article 35 of the GDPR, but also in national laws outside of Europe, such as, for instance, article 33 of the Personal Information Protection Act of South Korea,²⁸ and article 38 of the General Law on Personal Data Protection of Brazil.²⁹

DPIA is directly relevant for AI, at least in two respects. First, article 35(1) of the GDPR requires to perform DPIA when ‘a type of processing in particular using new technologies … is likely to result in high risk to the rights and freedom of natural subjects’. National data protection authorities such as the Italian³⁰ and British³¹ ones have confirmed that AI qualifies as innovative technology, triggering the obligation to perform a DPIA. Second, according to article 35(3) of the GDPR, a DPIA is required when there is a ‘systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing’. The term ‘automated processing’ is key and suggests situations where data processing could take place without human supervision, as the result of pre-defined conditions (eg if X, then Y), but also as the result of independent actions taken by the system through which the data processing occurs.³²

²⁵ See David Wright and Paul De Hert, ‘Introduction to Privacy Impact Assessment’ in David Wright and Paul De Hert (eds), *Privacy Impact Assessment* (Springer 2012) 8–9. See also Dariusz Kloza, ‘The Concept of Impact Assessment’ in J Peter Burgess and Dariusz Kloza (eds), *Border Control and New Technologies. Addressing Integrated Impact Assessment* (Uitgeverij ASP 2021) 33–34.

²⁶ Although the present contribution adopts the term Data Protection Impact Assessment (DPIA), the authors are aware of the alternative term of Privacy Impact Assessment (PIA) and of the scholarly debate about the differences of these terms. See eg Paul De Hert, ‘A Human Rights Perspective on Privacy and Data Protection Impact Assessments’ in David Wright and Paul De Hert (eds), *Privacy Impact Assessment* (Springer 2012) 33. See also David Wright and Charles Raab, ‘Privacy Principles, Risks and Harms’ (2014) 28 International Review of Law, Computers and Technology 277.

²⁷ Council of Europe, ‘Convention 108+. Protocol Amending the Convention for the Protection of Individuals with Regard to the Processing of Personal Data’, Council of Europe Treaty Series No 223, (2018), <https://rm.coe.int/16808ac918>.

²⁸ ‘Personal Information Protection Act’, National Assembly of the Republic of Korea 2020, Act No 16930 <https://www.privacy.go.kr/eng/laws_view.do?nttId=8186&imgNo=3>.

²⁹ ‘Lei Geral de Proteção de Dados Pessoais’ (Presidência da República 2019) Lei no 13.709.

³⁰ ‘Elenco Delle Tipologie Di Trattamenti Soggetti al Requisito Di Una Valutazione d’impatto Sulla Protezione Dei Dati Ai Sensi Dell’art. 35, Comma 4, Del Regolamento (UE) no 2016/679’ (Garante per la protezione dei dati personali 2018).

³¹ Information Commissioner’s Office, ‘Examples of Processing “Likely to Result in High Risk”’ (ICO) <<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias/examples-of-processing-likely-to-result-in-high-risk/>>.

³² According to the European High Level Expert Group on Artificial Intelligence, AI systems ‘act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal [and] can also adapt their behaviour by analysing how the environment is affected by their previous actions’. See Independent High-Level Expert Group on Artificial Intelligence, ‘A Definition of AI: Main Capabilities and Disciplines. Definition Developed for the Purpose of the AI HLEG’s Deliverables’

The need to perform DPIA in the context of AI is also found in article 29(6) of the recently published European proposal for an Artificial Intelligence Act (AI Act).³³ The AI Act proposal sets out the obligations for concerned stakeholders that intend to create and commercialise AI. Despite obligations to perform human rights impact assessments currently being discussed at the European Parliament,³⁴ there is no requirement in the AI Act to perform impact assessment other than DPIA when applicable.³⁵ Supporting documentation accompanying the proposal confirms it. Accordingly, the obligation to perform an impact assessment was ‘discarded, because users of high-risk AI systems would normally be obliged to do a DPIA that already aims to protect a range of fundamental rights of natural persons and which could be interpreted broadly, so new regulatory obligation was considered unnecessary’.³⁶ This refers to article 35(1) of the GDPR, which requires DPIA when the risk is to the ‘rights and freedoms of natural persons’. By employing the generic terms of ‘rights’ and ‘freedoms’ instead of the specific one of ‘data protection rights’, and by employing the term ‘natural persons’ instead of ‘data subjects’, the GDPR makes this provision open-ended, suggesting a scope that goes beyond data protection. This is in line with interpretative guidance from the Article 29 Working Party (WP29, now European Data Protection Board), which states that rights ‘of the data subjects primarily concerns the right to privacy but may also involve other fundamental rights’.³⁷

It seems DPIA can already be considered a human rights impact assessment, and, consequently, that DPIA performed in the context of AI can be considered a HRIA-AI. As such, HRIA-AI is already mandated by data protection norms.

(European Commission 2019) 6. However, it is worth noticing that scholars have highlighted that it is possible for automated decision-making system not to amount to AI systems and, nevertheless, have a risk profile equivalent to or higher than that of AI systems. See Michelle Seng Ah Lee, ‘Defining the Scope of AI ADM System Risk Assessment’ in Eleni Kosta and Irene Kamara (eds), *Research Handbook on EU Data Protection Law* (Edward Elgar 2022) 405.

³³ European Commission (n 16).

³⁴ See eg Committee on the Internal Market and Consumer Protection and Committee on Civil Liberties, Justice and Home Affairs, ‘Draft Report on the Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act/AI Act) and Amending Certain Union Legislative Acts’ (2022) 60.

³⁵ The AI Act proposal also envisages conformity assessment for high-risk AI systems. However, as scholars highlighted, conformity assessment under the purview of the AI Act proposal is substantially different from impact assessment and has a different scope than assessing impact to human and fundamental rights. See Nikolaos Ioannidis and Olga Gkotsopoulou, ‘The Palimpsest of Conformity Assessment in the Proposed Artificial Intelligence Act: A Critical Exploration of Related Terminology’ (*European Law Blog*, 2 July 2021) <<https://europeanlawblog.eu/2021/07/02/the-palimpsest-of-conformity-assessment-in-the-proposed-artificial-intelligence-act-a-critical-exploration-of-related-terminology/>>.

³⁶ European Commission, ‘Commission Staff Working Document Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules of Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts’ SWD(2021) 84 final, Pt 1/2, 58–59.

³⁷ Article 29 Working Party, ‘Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks’ (2014) 4.

However, by anchoring the obligation to perform impact assessment within the data protection framework, the regulators excluded the possibility to mandate HRIA-AI when personal data processing is not involved. As per article 2(1) of the GDPR, this applies only in the context of processing of personal data. Thus, when a use of AI does not involve such processing, the GDPR does not apply³⁸ and, consequently, the obligation to perform DPIA is not applicable. This may lead to situations when natural persons are exposed to risks to their rights. For instance, there are predictive policing AI systems that assess the crime rate from a geospatial perspective and provide a scoring that informs law enforcement decisions.³⁹ These may make individuals subject to personal freedom violations such as arbitrary search even without their personal data being processed by the system.⁴⁰ Or there might be situations when the AI system governing the news feed of a social media can decide to show users only negative news, with potentially psychological consequences for users affected by depression or suicidality.⁴¹

As seen above, as of now there are no requirements for assessing human rights impacts of AI systems, unless a personal data processing occurs. In this case, the assessment takes place in the form of a DPIA. However, the lack of a formal requirement to perform the assessment, does not exclude *a priori* a role of DPIA even when such processing does not occur,⁴² especially, in the context of the AI Act. Indeed, scholars have investigated how DPIA can be an optimal solution to respond to the human oversight requirement for high-risk AI systems set by the AI Act proposal.⁴³ Accordingly, DPIA is to be seen as a governance mechanism to ensure transparency and explainability of the AI system and to help demonstrate compliance. As an overarching governance mechanism, DPIA should cover all the phases of the AI system life cycle, including phases where personal data processing might not occur, such as the design and development phases. Therefore, DPIA can be interpreted as an instrument that has value also outside the context of personal data processing. Also, some scholars suggested that DPIA might be a valid starting

³⁸ See eg Rosamunde van Brakel, 'How to Watch the Watchers? Democratic Oversight of Algorithmic Police Surveillance in Belgium' (2021) 19 *Surveillance & Society* 228.

³⁹ See eg Adelson Araujo Junior, 'A Predictive Policing Application to Support Patrol Planning in Smart Cities,' 2017 International Smart Cities Conference (ISC2) (IEEE 2017).

⁴⁰ Will Douglas Heaven, 'Predictive Policing Algorithms Are Racist. They Need to Be Dismantled' (*MIT Technology Review*, 17 July 2020) <www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/>.

⁴¹ See eg Marc Goodman, *Future Crimes: A Journey to the Dark Side of Technology—and How to Survive It* (Transworld 2015) 322.

⁴² Scholars have investigated the role of DPIA to safeguard individual rights also in other specific contexts, such as women's rights protection in smart cities. See eg Alessandra Calvi, 'Gender, Data Protection & the Smart City: Exploring the Role of DPIA in Achieving Equality Goals' (2022) 19 *European Journal of Spatial Development* 24.

⁴³ See Guillermo Lazcoz and Paul De Hert, 'Humans in the GDPR and AIA Governance of Automated and Algorithmic Systems. Essential Pre-Requisites Against Abdicating Responsibilities' (2022) Brussels Privacy Hub Working Paper 8.

point to model HRIA-AI.⁴⁴ Section 3 deconstructs DPIA into its building blocks to understand where changes should occur to make it fit the assessment of human rights impacts of AI.

3 Data Protection Impact Assessment in the GDPR

To understand DPIA we should first understand the general meaning of 'impact assessment'.⁴⁵ An impact assessment is distinct from risk assessment insofar as it focuses only on the consequences while a risk assessment looks at the 'effect of uncertainty on objectives ... expressed in terms of risk sources, potential events, their consequences and their likelihood'.⁴⁶ Thus, even if risk assessment also considers the consequences of an event, these are looked at in the context of other elements to calculate a risk score or level (see Figure 35.1).

Given these definitions, impact assessment can be understood in two ways: as a standalone process or as a sub-process of risk assessment with the goal to evaluate consequences as a component of risk.⁴⁷ One of the reasons to perform an impact assessment as a standalone process is to have a better visibility on the consequences of an event and come up with more effective mitigation plans.⁴⁸

To understand which kind of assessment DPIA is, we turn to the GDPR. Recital 90 states that 'a data protection impact assessment should be carried out by the controller ... in order to assess the particular likelihood and severity of the high risk.' This suggests that impact assessment is a synonym for risk assessment since it involves assessing the likelihood and severity,⁴⁹ the two elements that constitute

⁴⁴ See eg Margot E Kaminski and Gianclaudio Malgieri, 'Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations' (2021) 11 International Data Privacy Law 20. See also Stephanie Sheir, *Algorithmic Impact Assessments: Building a Systematic Framework of Accountability for Algorithmic Decision Making* (Institute for the Future of Work 2021) 7.

⁴⁵ The International Association for Impact Assessment defines it as 'the process of identifying the future consequences of a current or proposed action'. See International Association for Impact Assessment, *What Is Impact Assessment?* (International Association for Impact Assessment 2009). The International Organization for Standardization defines consequence as the 'outcome of an event ... affecting objectives'. See ISO 31000:2018 'Risk Management—Guidelines'.

⁴⁶ ISO 31000:2018 'Risk Management—Guidelines'.

⁴⁷ See eg Henrik Hassel and Alexander Cedergren, 'Integrating Risk Assessment and Business Impact Assessment in the Public Crisis Management Sector' (2021) 56 International Journal of Disaster Risk Reduction <<https://www.sciencedirect.com/science/article/pii/S2212420921001023>>.

⁴⁸ An example of this kind of impact assessment is Business Impact Analysis (BIA), which can be defined as a process that 'analyses the effects of a disruption on the organization [the outcome of which] is a statement and justification of business continuity priorities and requirements'. See ISO/TS 22317:2021 'Security and Resilience—Business Continuity Management Systems—Guidelines for Business Impact Analysis'. In the context of information system, BIA can be employed to analyse 'information system's requirements, functions, and interdependencies used to characterize system contingency requirements and priorities in the event of a significant disruption'. See Marianne Swanson and others, 'Contingency Planning Guide for Federal Information Systems' (National Institute of Standards and Technology 2010) 800–34.

⁴⁹ Although less common, 'severity' is sometimes used as a synonym for impact or consequence in risk management terminology. See eg 'Enterprise Risk Management—Key Definitions—Definition of

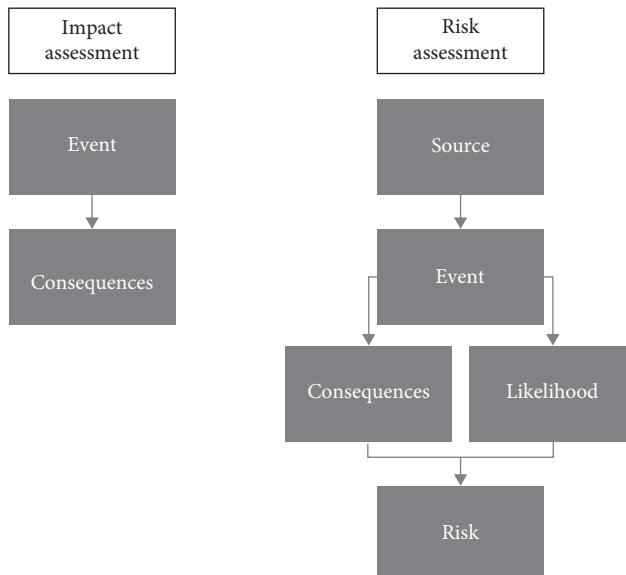


Figure 35.1 Comparison between impact assessment and risk assessment processes
Figure designed by the authors.

risk. Moreover, article 35(7)(c) of the GDPR provides that a DPIA ‘shall contain at least … an assessment of the risks to the rights and freedoms of data subjects’. In conjunction with Recital 90, the provision seems to suggest DPIA is not a standalone process separated by risk assessment, nor that DPIA is part of the risk assessment process. The GDPR seems to endorse a different approach from traditional risk management practices, according to which DPIA is a larger process that includes the assessment of risks within its boundaries.

Despite this terminological peculiarity, the GDPR provides great flexibility to data controllers. Article 35(7) of the GDPR lists the elements that the DPIA shall contain: ‘(a) a systematic description of the envisaged processing operations and the purposes of the processing …; (b) an assessment of the necessity and proportionality of the processing operations in relation to the purposes; (c) an assessment of the risks to the rights and freedoms of data subjects …; and (d) the measures envisaged to address the risks’. These are minimum requirements, and DPIAs can be scaled-up to include additional elements.⁵⁰ For instance, article 35(9) of the GDPR

Risk Severity’ (*Stanford University Office of the Chief Risk Officer*) <<https://ocro.stanford.edu/erm/key-definitions/definition-risk-severity>>.

⁵⁰ Raphaël Gellert, ‘Understanding the Notion of Risk in the General Data Protection Regulation’ (2018) 34 Computer Law & Security Review 279.

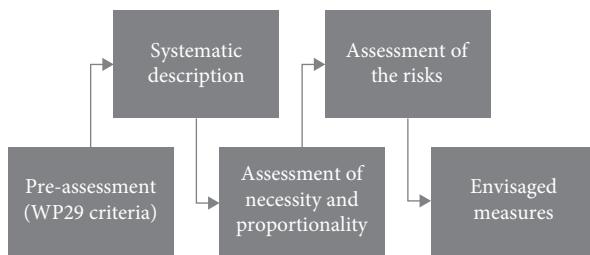


Figure 35.2 DPIA Process

Figure designed by the authors.

envisages the possibility to ‘seek the views of data subjects or their representatives’ where deemed appropriate.⁵¹

Article 35(3) of the GDPR lists cases when a processing is likely to amount to high-risk. However, this is not to be considered exhaustive, as also stressed by the WP29.⁵² To help the practitioners in the field, the WP29 drafted a list of nine criteria that can be used to evaluate if a processing is likely to result in high-risk. This exercise of checking the WP29 criteria should not be confused with a risk assessment, which involves the calculation of a risk score based on the likelihood and consequences of an event. The WP29 approach is a simple ‘ticking-the-box’ exercise against the nine criteria and represents a ‘pre-assessment’ to understand if a DPIA is necessary.

Looking at the blackletter rules of the GDPR, the DPIA can be deconstructed in five mandatory steps (see Figure 35.2). First, a pre-assessment to verify if a data processing amounts to high risk; second, a systematic description of the processing including of its purposes; third, an assessment of the necessity and proportionality of the processing; fourth, an assessment of the risks to the rights and freedoms of data subjects connected to the processing; and fifth, a description of the measures the data controller has or would put in place to mitigate the risks.

In addition to these five mandatory steps, there are at least two other elements to be considered when performing a DPIA. The first one is about the involvement of other interested parties in the assessment process, namely the Data Protection Officer (DPO)⁵³ as per article 35(2), the data subjects as per article 35(9), and

⁵¹ Scholars have highlighted how this specific, yet optional step is particularly suited to manage risks in the context of AI. See eg Ronan Hamon, ‘Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making’ (2022) 17 IEEE Computational Intelligence Magazine 72.

⁵² Article 29 Data Protection Working Party, ‘Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is “Likely to Result in a High Risk” for the Purposes of Regulation 2016/679’, 9.

⁵³ A thorough description of the role of DPO can be found in Article 29 Working Party, ‘Guidelines on Data Protection Officers’ (2017).

relevant supervisory authority as per article 36.⁵⁴ The second element to be considered is the process of reviewing the assessment. Article 35(11) of the GDPR requires such process to be performed ‘where necessary [and] at least when there is a change of the risk represented in the processing operations’. Monitoring the situation to understand when a review is necessary implies that the DPIA should be interpreted as a continuous exercise rather than a point-in-time activity.⁵⁵

4 From DPIA to HRIA-AI

In the following sections, we will take a closer look at the five steps mentioned in section 3 and how they can be customised to respond to the needs of a HRIA-AI and the specific challenges at hand—changing the technological landscape and human rights frameworks.

4.1 First Step: Pre-Assessment

The first step is what we termed pre-assessment. In the context of a DPIA, the data controller is required to run the pre-assessment to understand if a specific processing of personal data is likely to result in high risks to the rights of natural persons. Much like the GDPR requires a DPIA in case of high-risk processing, by analogy we can model a HRIA-AI to be necessary in case of high-risk AI systems. However, as highlighted above, AI does not necessarily involve the processing of personal data. Therefore, since—in such cases—the data protection framework would not apply, the criteria to understand whether a HRIA-AI should be performed should not be searched for in such framework.

One way to address this point is to look at the AI Act proposal. This adopts a risk-based approach that classifies AI systems in prohibited AI practices, high-risk AI systems, and low or minimal-risk AI systems.⁵⁶ According to article 6 of the AI Act proposal, AI systems shall be considered high-risk where they fall under the scope of one of the legal acts listed in Annex II to the proposal, or where they are under the scope of Annex III to the proposal.

⁵⁴ While the involvement of the DPO and of the supervisory authority is required under certain circumstances, the involvement of data subjects is optional. However, researchers have investigated the role of data subjects in the DPIA process. See eg Athena Christofi, ‘Data Protection, Control and Participation beyond Consent. “Seeking the Views” of Data Subjects in Data Protection Impact Assessments’ in Eleni Kosta, Ronald Leenes, and Irene Kamara (eds), *Research Handbook on EU Data Protection Law* (Edward Elgar 2022) 503. See also note 52.

⁵⁵ This position is confirmed also by the WP29. See Article 29 Data Protection Working Party (n 53) 16.

⁵⁶ It is interesting to notice how the European Parliament and the European Commission put forward different views on how to classify AI systems. See Inês de Matos Pinto, ‘The Draft AI Act: A Success Story of Strengthening Parliament’s Right of Legislative Initiative?’ (2021) 22 ERA Forum 619, 632–33.

Annex III provides a list of specific cases of high-risk AI systems.⁵⁷ As such, it can be considered a useful guideline to identify when a HRIA-AI should be performed. Think for instance to an AI system designated to elaborate the criteria for the creditworthiness of people. This specific AI system does not directly look at the individuals and, thus, it does not process personal data. The system only considers the general context and societal and economic circumstances to craft the criteria that can then be applied by human operators or other systems. For example, the AI system might conclude that people that work in a specific sector, that were first hired in a specific timeframe, and that work in a specific geography, are more prone to lose their job in the next 6 months than other people. The use of this system can potentially impact their access to loans or essential services such as bank accounts, even without any processing or personal data occurring. However, one of the criteria found in the AI Act proposal for classifying an AI system as high-risk is that AI system is employed ‘to evaluate the creditworthiness of natural persons or establish their credit score’. Even though the system does not directly evaluate natural persons, its intended use can effectively introduce risks for them. Therefore, while this situation would not fall under the scope of a DPIA, it might fall under the scope of a HRIA-AI that uses the criteria for high-risk system from the AI Act.

4.2 Second Step: Systematic Description of AI System

The second step of a DPIA is about providing a systematic description of the processing operations. Article 35(7) of the GDPR requires the description to include at minimum the purposes of the processing and the pursued legitimate interest. National data protection authorities have provided DPIA templates that go beyond these limited requirements.⁵⁸ Although each template adopts a different structure and terminology, all of them include the ‘nature’, ‘scope’, and ‘context’ of data processing among the aspects to be described.

In the context of a HRIA-AI, the assessment would focus on the systematic description of the AI system. Starting from this, concerned stakeholders can identify the most appropriate questions or approach to describe the nature, scope, context,

⁵⁷ Annex III presents six macro areas. These are: ‘Biometric identification’, ‘Management and operation of critical infrastructure’, ‘Education and vocational training’, ‘Employment, workers management and access to self-employment’, ‘Access to and enjoyment of essential private services and public services and benefits’, and ‘Law enforcement’. Each macro area is then detailed in one or more specific use cases.

⁵⁸ See eg Information Commissioner’s Office, ‘Sample DPIA Template’ (ICO, 2018) <<https://ico.org.uk/media/about-the-ico/consultations/2258461/dpia-template-v04-post-comms-review-20180308.pdf>>. See also Commission Nationale Informatique et Libertés, ‘Privacy Impact Assessment’ (2018) <<https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-2-en-templates.pdf>>; and Agencia española protección datos, ‘Gestión Del Riesgo y Evaluación de Impacto En Tratamientos de Datos Personales’ (2021) <<https://www.aepd.es/es/documento/gestion-riesgo-y-evaluacion-impacto-en-tratamientos-datos-personales.pdf>>.

Aspect	Question for concerned stakeholders
Nature of AI system	1) What is the source of the data fed into the algorithm? 2) How are data fed into algorithm (supervised v unsupervised)? 3) What underlying technology/characteristics is the AI system equipped with? (e.g., neural network, deep learning, federate learning, etc.)
Scope of AI system	4) What is the nature of the data fed in/generated by the algorithm (e.g., personal, non-personal, synthetic, etc.) 5) Who are the expected users of the AI system? 6) In what context is the AI system going to be implemented? 7) What users' rights will likely be impacted by the use (or misuse) of the AI system?
Context of AI system	8) What upstream/downstream dependencies exist for the AI system? 9) Is the AI system mission critical for your organisation?
Purpose of AI system	10) What goal does the AI system aim to achieve? 11) What goals could the AI system potentially satisfy in the future?

Figure 35.3 Proposal for a list of minimum requirements to be included in the systematic description of the AI system during HRIA-AI

Figure designed by the authors.

and purpose of the AI system. Figure 35.3 provides a non-exhaustive and illustrative list of questions to achieve this goal.

Description on the nature of the AI system can shed light on specific risks for the system and constraints that concerned stakeholders are likely to face when mitigating such risks. For instance, understanding if the AI system relies on supervised or unsupervised learning⁵⁹ can inform concerned stakeholders on the chance for the system to be affected by bias. The case of the Twitter chatbot, 'Tay', is illustrative. Released by Microsoft in 2016, Tay was shut down less than twenty-four hours from its deployment after it developed the tendency to publish discriminatory and offensive messages. The chatbot relied on unsupervised learning techniques to learn from Twitter interactions.⁶⁰ The lack of (effective) human oversight left the bot exposed to a targeted campaign from users of the well known bulletin board,

⁵⁹ Supervised and unsupervised learning refer to two approaches to train AI systems on how to make optimal decisions given a set of data. The former employs labelled data to supervise the training of the system, while the latter relies on self-discover techniques the system can directly put into action. See Richard Ribon Fletcher, Audace Nakashima, and Olusubomi Olubeko, 'Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health' (2021) 3 Frontiers in Artificial Intelligence 561802.

⁶⁰ Mark van Rijmenam, Jochen Schweitzer, and Mary-Anne Williams, 'Overcoming Principal-Agent Problems When Dealing with Artificial Agents: Lessons for Governance from a Conversation with Tay' (*ResearchGate*, May 2019) <www.researchgate.net/publication/333079462_Overcoming_principal-agent_problems_when_dealing_with_artificial_agents_Lessons_for_governance_from_a_conversation_with_Tay>.

'4Chan', who inundated Tay with offensive content which the system understood as the common parlance and used to build its own way of communicating.⁶¹

Description on the scope of the AI system can shed light on which human rights are likely to be impacted by the AI system. Without such an assessment, concerned stakeholders would be required to assess the impacts for every single right and freedom whenever they perform a HRIA-AI. Such an approach could make the process costly and possibly unsustainable. Thus, due to the excessive effort required, organisations might decide not to perform HRIA-AI and limit their approach to compliance with existing requirements (eg performing DPIA in the context of certain data processing). However, performing HRIA-AI on a subset of human rights could make the process easier and promote this voluntary exercise.

In some cases, identifying what rights to include in a HRIA-AI is relatively simple. For instance, we can reasonably assume that an AI system that displays selected advertisement does not affect the right of natural persons not to be subject to torture as per article 4 of the EU Charter of Fundamental Rights. Other times, identifying the relevant human rights is more difficult. For instance, in the application of an AI system just described, one might question if such an advertisement approach could constitute a violation of freedom of thought enshrined in article 10 of the EU Charter of Fundamental Rights.

An additional complexity for concerned stakeholders is to identify the full list of human rights from which to select the ones that are relevant to the HRIA-IA. Yam suggests that such list should include only human rights that are codified by the law.⁶² In addition to that, as Quinn⁶³ and Malgieri⁶⁴ demonstrated, there are situations that, despite not formally amounting to human rights violation, can nevertheless be detrimental for natural persons. Although we concur with this position, identifying such situations might be complex for concerned stakeholders. Thus, we suggest that organisations with fewer resources or lower maturity in the field of human rights, follow the approach suggested by Yam and focus on assessing AI systems versus codified human rights. The assessment of situations not codified by the law should be limited to organisations with more mature human rights due diligence processes.⁶⁵ This also includes any assessment exercise that aims to foresee or anticipate potential changes that might occur to the human rights framework

⁶¹ Oscar Schwartz, 'In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation: The Bot Learned Language from People on Twitter—But It Also Learned Values' (*IEEE Spectrum*, 2019) <<https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>>.

⁶² Yam and Skorburg (n 7) 9–10.

⁶³ Quinn (n 12).

⁶⁴ Gianclaudio Malgieri and Jędrzej Niklas, 'Vulnerable Data Subjects' (2020) 37 Computer Law & Security Review 105415.

⁶⁵ Scholars have also proposed to consider clusters of human rights and move from these aggregated sets to identify relevant ones. See Heleen L Janssen, 'An Approach for a Fundamental Rights Impact Assessment to Automated Decision-Making' (2020) 10 International Data Privacy Law 76–106.

(such as the emergence of new human rights). For instance, resourceful organisations might adopt foresight techniques to gain visibility into possible futures (such as anticipating scenarios that might spur from cultural and societal trends),⁶⁶ while less resourceful ones might limit their analysis to more immediate development (such as looking into existing proposals for new regulation).

Description on the context of the AI system can shed light on how impacts extend beyond the specific AI system under analysis. Think for instance to a kiosk placed inside an office and equipped with biometric recognition technology. Employees use the kiosk to clock in. When they do this the kiosk recognises the employees through their facial features. The kiosk then sends to the smartphones of the employees a temporary QR ('Quick Response') code they can use to open the areas of the building which they are authorised to access. However, the system is affected by an underlying bias that causes it to consistently mismatch people with certain features (eg the colour of their skin) and to give them wrong access permissions. In such case, the detrimental effects from the AI system of the kiosk are not immediate but manifest downstream when the employees present the QR code to the electronic lock of the areas they want to enter and are denied access.

Lastly, the description on the purpose of the AI system can be seen as an opportunity to adopt a forward-looking perspective into the evolution of the technology landscape. Indeed, purposes might evolve together with technology. The AI system might gain functionalities that make it possible and desirable for the organisation to expand its purpose. Although these new functionalities might require a review of the HRIA-AI once the decision to implement a new technology will be taken, capturing this expectation in advance can help to streamline the process. Moreover, suggested amendments to the AI Act proposal welcome this forward-looking perspective. In their draft report, the AI Act rapporteurs advise the Commission to consider the reasonably foreseeable misuses of an AI system when assessing if it should be considered high-risk.⁶⁷

Assessing what constitutes a reasonably foreseeable (mis)use is a case-by-case exercise. Let us consider the following example. An appliances manufacturer produces smart fridges. These are equipped with an AI system that can detect when the fridge is getting empty. The fridge cannot discern specific items. It performs such task by measuring the weight and volume of the items and cross-referencing the data with the consumption habits of the householders. When conditions are met, the fridge displays an alert on a dedicated notification screen. In this case, the purpose of the AI system could be described as to help users to better plan grocery shopping and improve consumption habits. However, the manufacturer develops

⁶⁶ For an overview of this field and its methods, see Wendell Bell, *Foundation of Future Studies. History, Purposes, and Knowledge. Human Science for a New Era: Volume 1* (5th edn, Transaction 2009).

⁶⁷ Committee on the Internal Market and Consumer Protection and Committee on Civil Liberties, Justice and Home Affairs (n 34).

a technology that allows the system to analyse and recognise single items stored in the fridge and considers implementing it in the next model to be put on the market. In the example, concerned stakeholders should assess the reasonably foreseeable (mis)uses of the new capability. This may vary according to the purpose behind its implementation. For instance, the fridge could simply use this capability to inform users of the items they need to pick up next time they go for grocery shopping. In this scenario, the impact to human rights is likely mild to non-existent. However, the same technology can be implemented in more invasive ways. The fridge could identify items with high sugar content and make this information available to third parties. In certain cases, such as for diabetic persons, this would have immediate consequences as insurance companies might obtain the information and decide to increase the cost of their health insurance.

Much like already described above for the analysis of impacted human rights, the extent of this forward-looking approach can vary according to the resources and capabilities available to the organisation and concerned stakeholders, ranging from foresight exercises (such as capturing weak signals of early-stage emergence of new technology and putting them into context with sociocultural dynamics),⁶⁸ to more limited assessments (such as the analysis of consolidating yet already emerged technology).

4.3 Third Step: Assessing Necessity and Proportionality

The third step is about performing the necessity and the proportionality tests for the envisaged use of the AI system versus the intended purpose of such use. Necessity can be defined as the criterion against which to assess the effectiveness of the measure in achieving intended objectives compared to other options that might be available for achieving the same goal.⁶⁹ Proportionality is a general principle according to which ‘the advantages resulting from the measure should not be outweighed by the disadvantages the measure causes with respect to the exercise of fundamental rights’.⁷⁰

The assessment of proportionality follows the assessment of necessity. Indeed, if an envisaged application or use of AI is assessed as being unnecessary, this should be dropped, and implementation should cease. However, the European Data

⁶⁸ For an overview of ‘weak signal’ as a foresight technique, see Paul JH Schoemaker and George S Day, ‘How to Make Sense of Weak Signals’ (*MIT Sloan Management Review*, 2009) <<https://sloanrev.ew.mit.edu/article/how-to-make-sense-of-weak-signals>>. See also Elina Hiultunen, ‘Good Sources of Weak Signals: A Global Study of Where Futurists Look For Weak Signals’ (2008) 12 *Journal of Future Studies* 21.

⁶⁹ European Data Protection Supervisor (EDPS), ‘Assessing the Necessity of Measures That Limit the Fundamental Right to the Protection of Personal Data: A Toolkit’ (2017) 5.

⁷⁰ EDPS, ‘EDPS Guidelines on Assessing the Proportionality of Measures That Limit the Fundamental Rights to Privacy and to the Protection of Personal Data’ (2019) 8.

Protection Supervisor (EDPS) cautions that this logic sequence must not be accepted strictly, since ‘there is some overlap between the notions of necessity and proportionality, and depending on the measure in question the two tests may be carried out concurrently or even in reverse order’.⁷¹

To understand the necessity and proportionality tests, let us get back to the example of the geospatial predictive policing tool mentioned in section 2. As already said, such tool can be used to inform the strategy of law enforcement, and certain strategies might result in a contraction of human rights of individuals. Law enforcement decides to adopt such a tool in response to a surge in criminal activities. Before implementing it, they perform the necessity test to answer the following question: does the situation require this predictive policing tool to contrast the surge in criminal activities, or are there other (less invasive) ways that can be employed? The assessor approaches this task by gathering data and information (eg through interviews with field officers) and concludes that, indeed, law enforcement would not be capable to manage this unusual surge in criminal activity without the support of the predictive policing tool. Thus, the assessor moves on to test the proportionality of the system.

To assess proportionality, the assessor compares the impacts coming from the surge in criminal activity and the impacts of using the predictive policing tools to answer the following question: are the impacts of the predictive policing tool justified and proportionate to the gravity of the situation at hand? By looking at the data, the assessor identifies that most of the criminal activities can be classified as minor offences, or misdemeanours, such as drug possession and shoplifting. The assessor then concludes that employing an invasive predictive policing tool is disproportionate compared to the impact of the crimes the police is trying to contain by using such a tool. Obviously, in case of more serious crimes such as physical assaults, considerations would be different.

A particularly challenging point of the necessity and proportionality tests is how to interpret them from an ethical standpoint. The interpretation of the necessity and proportionality tests can vary and is connected to the social and cultural contexts where these are performed. For instance, in the example above, the underlying assumption is that safeguarding the interest to crime prevention and control would inevitably weaken human rights of individuals. Thus, the necessity and proportionality test should be seen as exercises to strike a balance between these two competing interests. However, these tests can also be looked at from a different point of view, one that considers only where to put the threshold of what is acceptable in terms of erosion of fundamental rights, irrespective of the other interests at stake.⁷² Within this meaning, a measure that is more invasive than the identified

⁷¹ EDPS (n 70) 5.

⁷² De Hert (n 26) 54–55.

threshold would fail the test even if necessary and proportionate in the context of a balancing exercise.

4.4 Fourth Step: Assessing Risks to Rights

The fourth step is about assessing the risks for the rights of natural persons. Similar to the necessity and proportionality tests, risk assessment can be looked at in terms of execution (ie the approach or methodology to adopt), but also through the lenses of ethical interpretation. As for execution, there is no single way on how to perform a risk assessment. In recent years scholars have put forward proposals on how to implement this in the context of human rights and AI. Mantelero and Esposito, for instance, designed an evidence-based methodology that can be adopted by concerned stakeholders, supervisory authorities, and auditing bodies to monitor the risk management practices of organisations.⁷³

As for the ethical interpretation, this concerns primarily the aspect of risk acceptance. Where should organisations put the threshold for risk acceptance? What is the maximum risk score or risk level that they are willing to tolerate? In the context of corporate risk management, this would be (relatively) easy, and the decision to accept risks usually follows economic considerations. However, in the context of a HRIA-AI, the organisation is assessing the risks from the perspective of the users. Thus, questions arise: why should an organisation be entitled to decide the risk tolerance threshold of individual persons? Should not this decision be taken by the interested subjects? Again, considerations can fluctuate between a consequentialist approach and a categorical approach. The first aims to balance the benefits coming from a specific use of AI with the risks originating from it. The second identifies inalienable human rights and considers any use of AI that impact such rights as impermissible, despite their risk score or level. As stated by Yeung and Bygrave, a ‘violation of a fundamental right is a serious moral wrong, *tout court*’.⁷⁴

The issue here is not about what is practically possible to achieve with a risk assessment. Although human rights can be considered intangible elements that escape quantification, businesses and organisations routinely quantify all sorts of intangibles, and there are techniques to bring measurability to intangible

⁷³ Alessandro Mantelero and Maria Samantha Esposito, ‘An Evidence-Based Methodology for Human Rights Impact Assessment (HRIA) in the Development of AI Data-Intensive Systems’ (2021) 41 *Computer Law & Security Review* 105561. Mantelero has recently published an update to his research. See Alessandro Mantelero, *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI* (Springer 2022).

⁷⁴ Karen Yeung and Lee A Bygrave, ‘Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship’ (2021) 16 *Regulation & Governance* 10.

elements.⁷⁵ The issue, which goes beyond the scope of this contribution, is whether human rights should be quantified at all, or if they are to be categorised as moral absolutes to be safeguarded at all costs and in every circumstance.

Despite the fundamentally different approaches described above, there are practical steps that concerned stakeholders can take to address the issue. For instance, scholars and experts have advocated for more open and participatory approaches to risk management built on the provision found in article 35(9) of the GDPR.⁷⁶ Accordingly, organisations should seek the view of users and take decisions following such consultation. The approach ensures that concerned stakeholders can adopt decisions—including whether to adopt a consequentialist or categorical approach to risk assessment—that consider the position of involved persons.

4.5 Fifth Step: Mitigation Measures

Finally, the fifth step is about describing the envisaged mitigation measures. Mitigation measures could be both technical and organisational. The description should clearly outline how much and in what way the mitigation measures contribute to lower the risks for the human rights that have been identified and assessed in the previous steps.

The five steps described above constitute the core elements of our proposal for HRIA-AI. Much like already said for DPIA, these are to be interpreted as minimum elements, and concerned stakeholders are welcome to add further steps.

5 Conclusion: HRIA-AI Is a Baseline Approach That Requires Further Refining

The present contribution began by acknowledging how the evolving technological landscape and the changing human rights framework can make it difficult for concerned stakeholders to meaningfully address impacts to human rights in the context of AI. In section 2, we mentioned different impact assessment practices and identified Data Protection Impact Assessment as the most valuable model to create a specific voluntary HRIA-AI approach. Indeed, we maintained that in certain instances—that is, when the use of AI systems involves the processing of personal data—DPIA already amounts to HRIA-AI. However, being applicable only in the context of personal data processing, it falls short of addressing human rights impacts when personal data processing does not occur. Thus, we recognised the

⁷⁵ See Douglas W Hubbard, *How to Measure Anything. Finding the Value of ‘Intangibles’ in Business* (Wiley 2014).

⁷⁶ Christofi (n 54).

value of DPIA as a model to design a HRIA-AI that can be adopted also outside of personal data processing. In section 3, we looked at DPIA, to understand its main characteristics. In section 4, we put forward our proposal for a HRIA-AI, strictly following the structure of DPIA described in section 3. We considered the structure of the DPIA as presented by the GDPR and interpretative guidance from the WP29 and identified its building blocks. We then went through each of these building blocks in sections 4.1–4.5, customising it to satisfy the specific needs of AI and to the specific challenges in the contexts of technology and human rights.

Our proposal for a HRIA-AI is not rooted in a legal requirement. Therefore, HRIA-AI remains a voluntary exercise. However, differently from a DPIA, HRIA-AI are not bound by specific contextual elements (namely, the occurrence of personal data processing). As such, we believe that, despite its voluntary nature, HRIA-AI can be an instrumental tool to increase the overall safeguards to individuals in a society that is growingly adopting AI technology.

Our HRIA-AI approach should not be considered as a conclusive answer to the issues at the intersection of human rights protection and AI. There are aspects that are not fully addressed by our HRIA-AI model, such as the identification and quantification of human rights impacts at different scale described by Krupy and McLeod Rogers.⁷⁷ Moreover, we believe that our proposal for a HRIA-AI can be enriched through further work that focuses on clearly and precisely outlining methods and standards on how each step can be implemented, and what capabilities are needed to do so. For instance, we touched upon the potential value of foresight techniques and their role in addressing changes to the technology and human rights landscape. These additional elements would inform a more practical design for a HRIA-AI.

⁷⁷ According to their study, harm to human rights can materialise at different scales, from micro—damage occurring from a regular yet limited use of AI through time such as the loss of self-awareness coming from over reliance on digital assistance—to societal ones. When harm materialise at such extreme scale, it is difficult to detect, and can escape assessment exercises. See Tetyana Krupiy and Jacqueline McLeod Rogers, ‘Mapping Artificial Intelligence and Human Intersections; Why We Need New Perspectives on Harm and Governance in Human Rights’ in Aoife O’Donoghue, Ruth Houghton, and Seshauna Wheatle (eds), *Research Handbook on Global Governance* (Edward Elgar 2023).

Real-Life Experimentation with Artificial Intelligence

Elizaveta Gromova and Evert Stamhuis

1 Introduction

‘You cannot experiment with fundamental rights’ cries out one activist in the notorious *Netflix* documentary ‘Coded Bias’, protesting against experimental application of sensory data collection and processing with the use of algorithms. If one would take this stance to its final consequence, all but every real-life testing with innovative data technologies where personal safety, freedom, privacy, property, or self-determination is involved would have to be banned in every jurisdiction that adheres to respecting human rights. Where uncertainty is inherent to real-life testing, pressure on fundamental rights cannot be totally pre-empted in experimental situations. So, some risk-taking will always be part of the testing game. In an absolute risk-avoidance scenario, the road from the lab to real life cannot pass through a test phase. Consequently, the release of new technologies would be much more abrupt—and in effect not at all safer for individuals’ rights. We find that an unattractive scenario for new technologies, such as artificial intelligence (AI)-powered systems, whatever one’s assumptions are with regard to the promotion of innovation by permissive regulations.¹ Since much good can come from careful experimentation, we aim to make a case for a legally acceptable level of risk-taking.² We explore in this chapter how operators and regulators could entertain experimentation with appropriate counterbalance vis-à-vis human rights. These careful experimentation practices comprise in our view the precautionary consideration of potential human rights violations. Such caution will work out beneficial, not only to those persons whose rights may be on the line. When a risk materialises in an actual violation of rights, the operator of the experiment and/or the authority that allowed the experiment will be held accountable. Consequently, designing the

¹ Anna Butenko and Pierre Larouche, ‘Regulation for Innovativeness or Regulation of Innovation?’ (2015) 7(1) *Law, Innovation and Technology* 52–82; Christina Poncibo and Laura Zoboli, ‘Sandboxes and Consumer Protection: The European Perspective’ (2020) 8 *International Journal on Consumer Law and Practice* 1–22.

² For a similar point, see Marta Katarzyna Kołacz, Alberto Quintavalla, and Orlin Yalnazov, ‘Who Should Regulate Disruptive Technology?’ (2019) 10(1) *European Journal of Risk Regulation* 4.

right policy for justification or compensation from the outset would be in their interest as well.

Modern concepts of regulation and governance, such as smart regulation, good and agile governance, and responsive regulation suggest that regulation must be flexible, adapt to a rapidly changing world, and should be supportive to innovation.³ A regulatory sandbox is one of the tools with which regulators around the globe try to meet such standards. The term ‘sandbox’ originates from the technology sector, where a ‘sandbox’ is a closed virtual environment designed for safely isolated testing of a pre-market version of products or services.⁴ The players in the sandbox, the actors that take part in the experiment, experience the strengths and weaknesses of the product in its current state and are often invited to suggest improvements as co-creators in the process. The regulatory sandbox entails a temporary exceptional legal regime under supervision of a regulating authority. The real-life testing of a product (in the broadest sense) renders knowledge not only about the functionality of the product in the interaction between product and consumer, but also about the functionality of the existing regulatory regime. On the latter, the regulator may act to adapt regulatory instruments to the actual and future status quo of the product or market.

The agility of this regulatory approach fits very well with the state of information and communication technology. Fluidity is the hallmark, which leads to numerous pleas for flexibility on the side of the regulating authority. The literature shows that regulatory sandboxes gained their momentum in densely regulated sectors, especially financial markets, where it is evident that data-driven technology products take a large share in the practice of *in vivo* testing on those markets.⁵ For the latest generation of ICT, the Draft AI Act for the European Union (EU) comprises the establishment of regulatory sandboxes for AI-driven systems in articles 53–55, applicable across sectors.⁶

In this chapter, we aim to uncover the conditions under which allowing experimentation with AI-powered systems in predestined legal environments can be compatible with the requirements of fundamental rights protection. We will follow a doctrinal methodology in analysing the legal status quo, with a focus on the European Convention on Human Rights (ECHR), with its force in EU and non-EU states. We will study the exceptions allowed under the ECHR and the preventive

³ E Gromova and T Ivanc, ‘Regulatory Sandboxes (Experimental Legal Regimes for digital innovations) for BRICS’ (2020) 7(2) BRICS Law Journal 10–36.

⁴ A Stern, ‘Innovation under Regulatory Uncertainty: Evidence from Medical Technology’ (2018) Journal of Public Economics 145, 181–200.

⁵ Poncibo and Zoboli (n 1); S Philipsen, EF Stadhuis, and WM de Jong, ‘Legal Enclaves as a Test Environment for Innovative Products: Toward Legally Resilient Experimentation Policies’ (2021) 15(4) Regulation & Governance 1128.

⁶ European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts’ (Artificial Intelligence Act/AI Act) COM(2021) 206 final, 2021/0106(COD) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>>.

action obligations, developed by the European Court of Human Rights (ECtHR) in Strasbourg. From that twofold analysis, we intend to present recommendations for a legally resilient sandbox practice for AI systems. We begin with a brief description of the policy reality of regulatory sandboxes in section 2 and then move on to the human rights concerns that are connected to sandboxing, particularly with AI-powered systems, in section 3. In section 4, the legal foundations are described from which to develop resilience conditions, which we then elaborate upon in section 5, in part from our observations of the practices in some countries. A conclusive summary follows in section 6. We are aware that the topic merits much more attention than we can give it in this chapter—because of its limited scope. Consequently, we chose not to elaborate on all the theoretical implications of our analysis and tried to be as helpful to policy practice as possible. Some suggestions for more in-depth research are included in section 6.

2 The Concept of Regulatory Sandbox and Its Impact on AI Development

A regulatory sandbox itself represents an experimental regime which allows business entities that take part in the experimentation to temporarily apply regulatory relaxations during testing of a new service or business model based on digital innovation. The regulatory sandbox was recognised in 2016 as a solution that allows the application of regulatory reliefs under current legislation to permit important experimenting for the new digital products.⁷ Regulatory sandboxes aim at encouraging innovations by allowing business entities to test technologies in a ‘safe’ environment due to the fact that the experimentation is always under the regulator’s control.

Firms can benefit from the sandbox because of the opportunity to find out whether a business model is attractive to consumers, or how a particular technology works in real-life conditions. It can also support in identifying consumer protection safeguards that can be built into new products and services.⁸ Regulators take advantage of these temporary regulatory relaxations by researching during the experiment whether current legislation is still appropriate or should be reconsidered. Consumers benefit from the introduction of new and potentially safer products, as regulatory sandboxes foster innovation and consumer choice in the long run.⁹

⁷ Philipsen, Stamhuis, and de Jong (n 5).

⁸ Financial Conduct Authority, ‘Regulatory Sandbox’ 27-03-2022, last update 14-04-2023 <www.fca.org.uk/firms/innovation/regulatory-sandbox>.

⁹ Madiega, Van De Pol, ‘Artificial Intelligence Act and Regulatory Sandboxes’ (European Parliamentary Research Service) June 2022, last update 07-06-2023 <[www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI\(2022\)733544_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf)>.

Regulatory sandboxes were first introduced in 2016 as part of a government initiative to support United Kingdom (UK) Fintech companies.¹⁰ The British Sandbox has encouraged innovation in more than 500 companies, and in more than 40 of them, it led to further regulation.¹¹ After the success of the first cohort, most firms acknowledged a further incentive to apply: getting the ‘badge of honour’ of being accepted in the sandbox and proving their business model in a live and regulated environment increased their credibility with both customers and investors.¹² The Financial Conduct Authority (FCA) and partners extended the sandbox into a global regulatory sandbox in 2020.¹³

There is evidence to suggest the sandboxing in the Fintech sector is a success in economic terms. Figures show that the implementation of a regulatory sandbox in the sphere of Fintech innovations has resulted in an increase of investments from \$1.8 billion to \$19 billion in the last five years. Market capitalisation of all cryptocurrencies has increased 1,578 per cent in the past twelve months from \$14 billion to \$237 billion.¹⁴ As these sources report, regulatory sandboxes’ experience led to the creation of more innovation-sensitive regulation and gave opportunity to the development of new business and deployment of new technologies. Several reports on regulatory sandboxes stated good performance on the indicators that were set. Most of the firms are continuing towards a wider market launch after the test in the sandbox. In addition, the majority of firms have gone on to secure their position on the financial market by applying for a full authorisation as financial provider.¹⁵

Regulatory sandboxes aimed at the development of AI exclusively have sprouted since 2020. Thus, the Norway Data Protection Authority launched a regulatory sandbox for responsible AI.¹⁶ Moreover, a five-year-long AI regulatory sandbox

¹⁰ Philipsen, Stamhuis, and de Jong (n 5).

¹¹ Zhang, Rowan, Duff, Homer, Schizas, Soriano, Cloud, Umer, Garvey, Ziegler, Wardrop, Blandin, Gray, Chen, Yerolemou, Calabria, Chantramonklasri, Dasgupta, Jenweeranon, Lin, Shuang ‘Early Lessons on Regulatory Innovation to Enable Inclusive FinTech: Innovation Offices, Regulatory Sandboxes and RegTech’, SSRN Electronic Journal, 30 June 2022, last update 07.06. 2023, ssrn.com/abstract=3621258

¹² Strachan, Nair, ‘A journey through the FCA regulatory sandbox: The benefits, challenges, and next steps’ (Deloitte), 2018, last update 07.06.2023, <www2.deloitte.com/uk/en/pages/financial-services/articles/journey-through-financial-conduct-authority-regulatory-sandbox.html?ysclid=lil287wdfr1919461912>.

¹³ The World Bank Brief, ‘Key Data from Regulatory Sandboxes across the Globe’ (World Bank), 1 November 2020, last update 07.06.2023, <www.worldbank.org/en/topic/fintech/brief/key-data-from-regulatory-sandboxes-across-the-globe/>.

¹⁴ World Bank, ‘Global Regulatory Sandbox Review’ Global Experiences from Regulatory Sandboxes. Fintech Note, No 8 (World Bank, Washington, DC), 11 November 2020, last updated 07.06.2023 <hdl.handle.net/10986/34789>.

¹⁵ Financial Conduct Authority, ‘Regulatory sandbox lessons learned report’, Financial Conduct Authority 2017, last updated 07.06.2023, <www.fca.org.uk/publications/research/regulatory-sandbox-lessons-learned-report>.

¹⁶ Pop, Adomavicius, ‘Sandboxes for Responsible Artificial Intelligence’, EIPA, September 2021, last update 07.06.2023, <www.activedemand-assets.s3.amazonaws.com/content_assets/6062/assets/EIPA_Briefing_Sandboxes_for_Responsible_Artificial_Intelligence.pdf?X-Amz-Expires=300&X-Amz-Date=20230607T021913Z&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credent>.

was launched in Moscow on 1 July 2020.¹⁷ In France, the CNIL Sandbox Initiative for Health Data was launched in 2021, covering innovative projects in the health-care sector that make use of personal data. The objective is to help organisations implement privacy by design from the outset.¹⁸ The UK was planning to launch a sandbox for Artificial Intelligence in Healthcare,¹⁹ while the same is true for Canada.²⁰ The Maltese Digital Innovation Authority also explicitly mentions experiments with AI in its communication of a technology assurance sandbox.²¹ Last but not least, the Draft AI Act of the European Commission, now under discussion among the various actors in the EU legislative process, provides for a regulatory sandbox for high-risk AI systems.²²

As summarised above, testing AI in regulatory sandboxes can help in improving regulation because of the possibility of evidence-based law-making. It can also promote the development of innovative AI solutions, help businesses to ensure compliance with relevant regulations, and create additional flexibility in terms of regulatory burden as well. But the discussion would be incomplete without dealing with concerns regarding the protection of rights of individuals. Such concerns have been voiced in the literature and regard not only basic consumer protection but also the right to privacy and personal data protection.²³ At the same time, those concerns incentivise *ex ante* testing of the systems in real-life circumstances, to discover and minimise human rights risks in a timely fashion.²⁴ We will take a closer look at these concerns in section 3.

¹⁷ AKIAXYJAIQN2QLVAGAOP/20230607/us-east-1/s3/aws4_request&X-Amz-SignedHeaders=host&X-Amz-Signature=31efc41bb9355c9ba263dda06ce00221e4e4170ec7c8e446c742c13f5ec4d3d0>.

¹⁸ TASS, 'Russia Develops a Regulatory Sandbox in the Field of Telemedicine', 8 February 2022, last update 07.06.2023, <www.ict.moscow/en/news/russia-develops-a-regulatory-sandbox-in-the-field-of-telemedicine/?ysclid=lil35qjvia159467241/>.

¹⁹ CNIL, 'France: CNIL selects digital health organisations for data protection sandbox', CNIL, 04 May 2021, last updated 07.06.2023, <<https://www.dataguidance.com/news/france-cnil-selects-digital-health-organisations-data>>. Commission Nationale de l'Informatique et des Libertes, Rapport d'activité 2020 (May 2021) 115<>.

²⁰ Downey 'Regulatory sandbox for AI needed to test and build systems', NHSX, 12 February 2020, last updated 07.06.2023, <www.digital.health.net/2020/02/regulatory-sandbox-for-ai-needed-to-test-and-build-systems-nhsx-says/>.

²¹ 'Canada is Planning to Launch RS in Healthcare' <www.bioworld.com/articles/498030-health-canada-making-use-of-regulatory-sandbox-to-address-ai?v=preview>.

²² Malta Digital Innovation Authority, 'Technology Assurance Sandbox' <<https://mdia.gov.mt/sandbox/>> undated.

²³ Artificial Intelligence Act (n 6) arts 53–55.

²⁴ A Stern, 'Innovation under Regulatory Uncertainty: Evidence from Medical Technology' (2018) 145 *Journal of Public Economics* 181–200; J Müller and A Kerényi, 'The Need for Trust and Ethics in the Digital Age: Sunshine and Shadows in the FinTech World' (2019) 3 *Financial and Economic Review* 5–34; Poncibo and Zoboli (n 1).

²⁵ Artificial Intelligence Act (n 6), Explanatory Memorandum, section 3.5.

3 Fundamental Rights Concerns About AI Sandboxes

The concerns or objections against the experimentation with AI in regulatory sandboxes are fed by two streams of critique against that practice. The first is related to the argument against regulatory sandboxing practices *per se*, that is, as far as sandboxing with any type of innovative technology is concerned. The creation of temporary law for specific actors would create unwarranted inequality and uncertainty, with disturbances of the level playing field, too much focus on accommodating business actors, and too little attention for consumers and other individuals.²⁵ The second critique is found in the hesitations and/or straightforward objections against the technology itself. Artificial intelligence should go through much more development, implementation research, and oversight before it could be safely permitted to have an impact on individual lives and society at large.²⁶ So, in this view, it is untimely—and even inappropriate—to create regulatory space for experimentation with this technology.

It is not our intention to reiterate these concerns or deal with them separately. They show overlap in the worries related to the position of individual consumers/citizens, where a negative impact on fundamental rights may flow from AI experimentation practices. We will address this common element by considering the fundamental rights that might be at peril in a practice of sandboxing with AI. We will then elaborate on the direction for dealing with this impact through preventing legal shortcomings in a regulated practice of sandboxing with AI. In other words: what legally inspired guidelines would be advisable to arrive at a trustworthy test practice with AI-driven technology applications?

For business actors regulated experimentation practices may also be legally problematic, for example, if they are unlawfully excluded from the sandbox.²⁷ However, the position of competing business actors on the relevant markets is outside the scope of this chapter. We have already clarified that fundamental rights of individuals are the focus of our attention. The prominent human rights instruments in the world award such rights to humans and not directly to corporations and companies.

There is a plethora of AI-driven systems for which one could wish real-life experimentation. As we have pointed at, there is an interest in current AI regulatory sandboxes in health applications, but it is not limited to that. With regard to the Draft EU Regulation for AI sandboxes there is also a great variety of applications that fall under article 54, concerning high-risk AI systems operating in many sectors.²⁸ Accordingly, it is not hard to imagine situations that pose risks

²⁵ Philipsen, Stadhuis, and de Jong (n 5).

²⁶ K LaGrandeur, ‘How Safe Is Our Reliance on AI, and Should We Regulate It?’ (2021) 1 AI Ethics 93–99.

²⁷ Philipsen, Stadhuis, and de Jong (n 5).

²⁸ Draft AI Act, art 6 and Annex III.

to fundamental rights of individuals. AI-driven systems can impact upon a great number of fundamental rights because the AI technology itself can be deployed in so many walks of life with so many divergent interests at stake.²⁹ For the EU, article 9 of the Draft AI Act postulates that an appropriate risk-management system should be in place for high-risk AI systems to be allowed on the EU market. In the development of the ‘continuous iterative process’ to ‘identify and analyse known and foreseeable risks’, as is required in article 9(2), the *in vivo* testing can be a crucial phase, where outside the lab new risks may arise.

The most salient example of testing may be the experimental allowance of fully autonomous vehicles on the public road network in interaction with other (uninformed) road users—on car lanes and crossings.³⁰ Life and limb of the persons in the car and of the other road users are clearly at stake in such experimentation. The right to life, as embedded in article 2 of the ECHR, immediately comes to mind. Also, the damage to other cars in case of a collision is relevant under Article 1 of the First Protocol (A1P1) to the ECHR (right to property). Other AI-powered products would carry other risks in the experimental situations, such as interferences with the private lives of individuals,³¹ in particular in cases where the experimental AI-driven system has the capacity of autonomous data collection and combination. The allowance of it to be tested poses risks to the rights enshrined in article 8 of the ECHR (right to private life). The pre-release testing of FinTech applications that are powered by AI we mention as another exemplary case. These services can cause financial harm for the consumer when they turn out to malfunction, which brings us back to the enjoyment of property (as safeguarded to a certain extent under A1P1).

As illustrated with the examples here above we fully admit that fundamental rights can be at risk in real life testing procedures. Notwithstanding, we still contend

²⁹ See Explanatory memorandum to the Draft AI Act, section 3.5: ‘The use of AI with its specific characteristics (eg opacity, complexity, dependency on data, autonomous behaviour) can adversely affect a number of fundamental rights enshrined in the EU Charter of Fundamental Rights (“the Charter”). This proposal seeks to ensure a high level of protection for those fundamental rights and aims to address various sources of risks through a clearly defined risk-based approach. With a set of requirements for trustworthy AI and proportionate obligations on all value chain participants, the proposal will enhance and promote the protection of the rights protected by the Charter: the right to human dignity (Article 1), respect for private life and protection of personal data (Articles 7 and 8), non-discrimination (Article 21), and equality between women and men (Article 23). It aims to prevent a chilling effect on the rights to freedom of expression (Article 11) and freedom of assembly (Article 12), to ensure protection of the right to an effective remedy and to a fair trial, the rights of defence and the presumption of innocence (Articles 47 and 48), as well as the general principle of good administration. Furthermore, as applicable in certain domains, the proposal will positively affect the rights of a number of special groups, such as the workers’ rights to fair and just working conditions (Article 31), a high level of consumer protection (Article 28), the rights of the child (Article 24) and the integration of persons with disabilities (Article 26). The right to a high level of environmental protection and the improvement of the quality of the environment (Article 37) is also relevant, including in relation to the health and safety of people.’

³⁰ Examples of AI potential in clinical practices are manifold, where the rigorous systems of certification of medical devices anyway require a high level of testing in clinical settings.

³¹ Poncibo and Zoboli (n 1) 17–18.

that an outright ban on experimentation is not the way forward. Therefore, we need to discover whether there is a middle ground, where a permissive policy for experimentation can be combined with due respect for fundamental rights. For that compromise, we will turn to the exception regimes that open avenues for justification of interferences with those rights. But there is more to study for discovering the full extent of legal obligations that push regulatory authorities towards a proactive policy vis-à-vis fundamental rights at peril in the sandbox practice.

It is not self-evident that allowing a rise in a certain risk for individuals' rights during an experiment constitutes a violation of that right on its own. So, when regulating authorities allow for a specific legal regime for testing AI-powered applications in real life, they do not *per se* already constitute a breach of human rights law (HRL).³² Creating or allowing a greater risk is not an interference with rights *tout court*³³ as a result of which some risk-taking does not activate the justification clauses. Another ground to build on would be the obligations to prevent interferences. Those exist, and we will briefly discuss them below, in order to explore whether a combination of these preventive obligations and the justifications regime for actual interferences can inspire us to propose guidelines for appropriate risk policies in regulatory sandboxing. We hope that this line of argumentation assists to absorb in concrete cases the legal debate on the question whether the risk leads to an interference and whether this interference constitutes a violation of the individuals' rights. Instead of that legal debate, we promote the proactive consideration of counterbalancing the risks for human rights at the moment that regulators design sandboxes and draft concrete conditions for experimentation.

For the EU member states, we can additionally point at the draft obligation in article 9(4) of the Draft AI Act. An AI-powered system shall have to comprise 'elimination or reduction of risks as far as possible through adequate design and development and, where appropriate, implementation of adequate mitigation and control measures in relation to risks that cannot be eliminated'. As expressed above, the testing phase can lead to the design and development reiterations that are implied here, but the testing itself needs to be accompanied as well with the appropriate measures to remedy or relieve the threats for individuals, even if the regulator allows temporary relief from compliance with article 9 during the sandbox period. One could say that the precautionary principle, implicit in article 9, also extends to the pre-release stage of the AI-powered system.³⁴

³² Florina Pop, 'Sandboxes for Responsible Artificial Intelligence' (*EIPA Briefing*, September 2021) <www.eipa.eu/publications/briefing/sandboxes-for-responsible-artificial-intelligence/>, states this with regard to data protection rights.

³³ We should mention that we do not refer to the 'significant disadvantage-test' (ECHR, art 35(3)). That test is indeed not about violations, but about access to the ECtHR. See DJ Harris and others, *Harris, O'Boyle, & Warbrick: Law of the European Convention on Human Rights* (3rd edn, OUP 2014) 68.

³⁴ The principle is explicit for environmental law (Treaty on the Functioning of the European Union (TFEU), art 191(2)) and relevant also for other policies, such as consumer protection. See Jale Tosun, 'How the EU Handles Uncertain Risks: Understanding the Role of the Precautionary Principle' (2013) 20 *Journal of European Public Policy* 1517–28.

4 Preventing the Dirt in the Sandbox

Abstaining from violations of fundamental rights should in principle be the duty of every legally relevant agent. The enforcement mechanisms in case of breaches of this obligation vary, according to the nature of the relation between the agent and the impacted person. Traditionally, human rights enforcement in relations between private persons relies on civil litigation, whereas protection systems such as the ECHR are open only to complaints from private persons against rights' violations by the state. This traditional distinction however does not represent the status quo, where the ECHR remedies are also open to complaints against the state for falling short in the protection of one person against violations by another private agent. This positive obligation of the state to protect is either explicit in the text of the ECHR (articles 2, 3, and 6) or read into it by the ECtHR in Strasbourg, as we will detail below.

As regards the normative framework for experimentation/sandboxes this state of the law means that private operators have a duty vis-à-vis the persons in and around the experiment to respect their rights.³⁵ In the given example of the AI-powered self-driving vehicle or vessel, we look at the passengers in the car or on board the ship and the ignorant passer-by on the road or waterway. The passengers one could ask for consent to the risks, the passer-by we cannot. No less important than the obligation of the operator is the duty of the state to deploy a reasonable effort to guard against fundamental rights violations between private agents in the experiment, in addition to its duties under the ECHR in case the state is the operator or co-operator itself.³⁶

The first column in the construction of risk prevention duties on the side of the authorities under HRL is found in the preventive obligations for the state. These are developed for life and health protection under article 2 of the ECHR, based on the seminal 1998 cases of *LCB v United Kingdom*³⁷ and *Osman v United Kingdom*,³⁸ but also for private and family life in the 1985 case of *X and Y v the Netherlands*,³⁹ with further reference to the judgment in *Airey v Ireland* from 1979.⁴⁰ A gradual development followed towards wider duties of the state for a larger group of rights.⁴¹ We will not deal with the full breadth of this case law, but only insofar as it accounts for precautionary duties in the context of allowing experimentation with AI systems.

³⁵ Under national tort law increased risk may violate duties of care and lead to litigation in case of serious risk for future damage, requiring an injunction. Article 53(4) of the Draft AI Act explicitly postulates civil liability for participants in the sandbox under Union and Member State law. We leave this out of our current study.

³⁶ As is shown in Philipsen, Stamhuis, and de Jong (n 5), the position of the state authorities in sandboxes is hardly ever as clearly distinct from the operator of the experiment as one would expect.

³⁷ *LCB v United Kingdom* App no 23413/94 (ECtHR, 9 June 1998).

³⁸ *Osman v United Kingdom* App no 23452/94 (ECtHR, 28 October 1998).

³⁹ *X and Y v the Netherlands* App no 8978/80 (ECtHR, 26 March 1985).

⁴⁰ *Airey v Ireland* App no 6289/73 (ECtHR, 9 October 1979).

⁴¹ Harris and others (n 33) 207–12, 504.

Essentially, the state needs to safeguard that real and immediate risks for human rights violations are reduced by appropriate measures, so that the actual breach of those rights does not occur in the probable scenarios of the test. The nature of those measures depends on the specifics of the risk, such as the chance of it realising, the seriousness of the impending breach, and the nature of the right under threat. For example, for risks of life and limb, other precautions are required than for a temporary limitation of the right to freedom of expression.⁴² A risk for property damage requires another compensation mechanism to be put in place than a risk for breaching informational privacy. We will detail examples of compensations, in part taken from actual practices, in section 5, but want to mention here that the deployment of precaution often materialises in a requirement of consent to the participation in an experiment, on the part of the individual whose position is under threat. For FinTech sandboxing, that may be a viable option because the operating firm can properly inform every consumer before participation in the experiment. For the ignorant passers-by in our AI-powered autonomous vehicle/vessel example, informed consent is more difficult to realise. The same will be true for, for instance, public safety and crime prevention use cases of AI, where the impacted persons are yet to be discovered. Those persons' rights should not be illusory, even when they are not directly in the visor of the operator or regulator of an experiment, as a result of which consent is not the panacea.⁴³

We now move to the second column in the construction for remedy and relief, derived from the justification clauses in the ECHR. Some of the rights under the ECHR are not absolute, because the Convention provides an exception regime for them. The second paragraphs of the articles 8–11 and of A1P1 provide the conditions under which a violation of the right(s) included in the first paragraph can be justified. There are slight differences between the formulations, but the similarity is obvious. One important collective aspect of this justification is the proportionality test. In its assessment of those conditions in concrete cases, the ECtHR has shown that a 'margin of appreciation' plays an important role. We will detail the various elements of the exception clause and conclude with the effect of the margin of appreciation in the remainder of this section.

First, the interference with the rights of individuals should be described by law.⁴⁴ This expresses the necessity of accessible and foreseeable allocation of the competence to act and the conditions for lawful action. As regards contemporary sandboxing practices, not all of them have such a clear legal fundament in

⁴² Eg while testing a moderation algorithm content may have been removed that in hindsight is not qualified as hate speech. For details on automated content moderation, see C Shenkman, D Thakur, and E Llansó, *Do You See What I See: Capabilities and Limits of Automated Multimedia Content Analysis* (CDT 2021).

⁴³ Philipsen, Stamhuis, and de Jong (n 5) put forward another consent-related problem: the tension between consent and scientific quality of testing outcomes, especially for testing responses in a random controlled trial setup.

⁴⁴ Harris and others (n 33) 506–09.

either statute or other sources of law. For example, the Fintech sandboxes rely for a large part on unspecified use of regulatory discretion on the side of the financial regulator. At the moment, for AI testing the national legislators need to have taken action. In the EU member states, as soon as the sandboxing regulation in the Draft AI Act enters into force, the Act will provide the description by law, probably complemented by appointment of the competent authority in member states' implementing legislation. For the regulators in the EU member states this future legislation will not only allocate the power to deploy regulatory sandboxes, but also set out more detailed conditions for legitimate sandboxing especially as regards the required levels of accessibility and foreseeability. For jurisdictions outside the EU, it is commendable that the relevant regulatory authority gets this 'described by law' criterion covered, even if only by publishing its own policy options and conditions for experimentation under a sandbox regime. This aims at preventing that within a certain jurisdiction a 'Wild West' of experimentation develops without openness to citizens and companies, leading to a race to the bottom when it comes to (costly) protection of interests other than those of the operator of the experiment. Others should be able to know who is to benefit and where and under what condition possible risks may arise.

Second, for the interference the pursuance of a legitimate aim is required. Listed in the text of the ECHR are the interests of national security, public safety or the economic well-being of the country, the prevention of disorder or crime, the protection of health or morals, or the protection of the rights and freedoms of others. As Harris and others observe, the objectives listed are phrased fairly widely, which renders it easy to comply.⁴⁵ The ECtHR is nevertheless strict in requiring an aim that is listed in the ECHR text,⁴⁶ so the authorities should give this due consideration in order to fall under the justification. The promotion of innovation and the development of new business models and technologies is not listed there, so the risk-taking in sandbox practices cannot be justified in isolation from other goals further downstream. One could argue that economic well-being is promoted by a positive climate for innovation, but that is hard to establish empirically in a given case. The breadth of AI applications opens the gateway to establish the connection to the ultimate objective of the tested system under one of the other legitimate aims. For the *ex ante* balancing of risks and remedies, we find it anyhow better to concentrate on the future benefits that the AI system to be tested may bring and then balance those with the relative weight of the impact on fundamental right by allowing the test in real life and specify the conditions for that event.⁴⁷

⁴⁵ Harris and others (n 33) 509–10.

⁴⁶ *SAS v France* App no 43835/11 (ECtHR, 1 July 2014), para 114.

⁴⁷ Convention 108+ (Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data), Council of Europe, CETS No 223, art 11(1)(a) provides for a rest category in the legitimate aims: 'other essential objectives of general public interest'. Whatever the weight of well-guided technological innovation might be, we still advise to put the ultimate goal, to be served by the AI

The third cumulative condition under which an interference with one of the rights of articles 8–11 of the ECHR can be justified is the criterion of ‘necessity in a democratic society’. It is held that this entails an extra test whether the interference is proportional to the objectives of the actions that constitute the interference and that there is a corresponding pressing social need.⁴⁸ Circumstances to observe are (i) the nature, effect, and period of the actions; (ii) their purpose; and (iii) the chances to achieve that purpose in the specific situation. As regards experimentation with AI systems, this confronts the regulatory authority with an extra burden to assess the means-end relation of allowing risk exposure in light of the rights’ violations that can occur and prepare for timely countermeasures when risks materialise in the given experiment.

To the letter, the exception-clauses discussed hereabove are only relevant, when there is a concrete interference. As pointed out, increased risk is not per se an interference with the right, and therefore a justificatory exception may not be directly applicable. In our perspective of proactively handling risks for an interference while designing and deploying a regulatory sandbox, the better representation is a spectrum of human rights’ impact, with completed rights violations on one side and a non-risk on the other side. With that in mind, the common principles of proportionality can be re-read as instructing to prevent, to counterbalance, and to offer remedies. Not all imaginary human rights impacts need to be addressed, but only those that are real and immediate. However, if the burden of counterbalancing measures is less heavy, the burden of the immediacy requirement could to a certain extent be relieved. And if the violation can have more serious consequences, the level of probability to consider would go down. So, the variety of potential impact is reflected in the variety of remedy and relief. This is not to say that unreal risks need to be considered as well, but in applying the principle of precaution the real but unknown risks are to be included. The likelihood of an impact can and should be included in the drafting process for the risk policy, an exercise that will take place anyway when insurance is sought for. A provisional discussion of various options for these measures will be presented in section 5, in part inspired by what we have seen in separate sandbox practices. We now have yet to deal with the ‘effective moderation’⁴⁹ that comes from the doctrine of ‘the margin of appreciation’. The careful observation of the margin of appreciation is relevant for both columns of the construction, outlined above, that is: for the preventive duties and for the measures necessary to fall under the justification clauses.

system, also in the scales, because that better incentivises the balancing to remain concrete and not a *pro forma* abstraction.

⁴⁸ Harris and others (n 33) 511.

⁴⁹ Dean Spielmann, ‘Whither the Margin of Appreciation?’ (2014) 67(1) Current Legal Problems 54.

With the doctrine of the margin of appreciation, the ECtHR applies an extra lever, restating the duties of the national authorities, to carefully perform the balancing act for all the interests at stake, with the ECtHR as a backstop mechanism. So, the ECtHR asserts its role as safeguard, recognises the stratification of enforcement mechanisms, and stresses the fair assessment of those interests in the concrete cases to take place at the level of national authorities and courts first.⁵⁰ Because the test of the margin of appreciation is not 'a fixed unit of legal measurement'⁵¹ its working matches very well the spectral nature of the risks for unlawful interferences at play in AI sandboxing. This doctrine should, however, not be interpreted as rendering the proportionality test in the stage of designing the sandbox rules a tepid affair. Although more than one precondition may be fit for purpose to pre-empt a risk for violations of human rights, the regulatory authority should take their duty seriously. So, the regulatory authorities can choose from a catalogue of relief and remedy mechanisms in their design of the AI sandbox, as long as they are proportional to the risk and do not make the protection of fundamental rights illusory. Some of these mechanisms we will exemplify in section 5.

5 Proactive Measures for Remedy and Relief

Combining the undeniable potential of interferences with fundamental rights (section 3) and the avenues for preventing a violation of the ECHR (section 4) leads us to the following conclusion. The authorities in the regulatory sandbox practices need to safeguard the legality of the real life experiments by applying a multifaceted risk policy. In that policy imperfections may exist, but need to be fixed by implementing the provisions based on the principles of credibility and trustworthiness.⁵² Human rights law mandates such risk measures when fundamental rights are under threat and at the same time allows for flexibility in ways to stay on the safe side of the law. That flexibility reflects the variety of risks and results in a toolkit full of proactive, repressive, and ex post measures to be undertaken. We mention but a few here, matching the focus of our chapter on ex ante instruction of the authorities who embark upon the actual activation of a regulatory sandbox practice.

The regulatory authority and the intended operator of the experiment can start the preparation of an appropriate risk strategy by applying one of the available

⁵⁰ See *ibid*, on the development of this doctrine since 1956 (ECommHR) and 1971 (ECtHR), culminating in *Handyside v United Kingdom* App no 5493/72 (ECtHR, 7 December 1976); see also Harris and others (n 33) 519.

⁵¹ Spielmann (n 49) 56.

⁵² AA Pinem and others, 'Trust and its Impact Towards Continuance of Use in Government-to-Business Online Service' (2018) 12(3/4) *Transforming Government: People, Process and Policy* 265–85.

human rights impact assessments.⁵³ In the risk policy of the sandbox, the concrete intervention must be classified in the risk spectrum we have laid out above and that instructs the nature and level of appropriate countermeasures. Broadly speaking, financial risks (against A1P1) correspond with financial indemnification to be available without cumbersome procedures. This translates into securing sufficient funds on the side of the operator and a low threshold procedure for establishing damage and compensation. Sufficient resources for indemnification can be found in either mandatory liability insurance on the side of the operator or warranty of indemnification from the state authority. The latter may appear too risky for the state, but we mention it explicitly because financial burdens on operators tend to benefit large incumbent firms, which is not necessarily good for the accessibility of the regulatory sandbox. The same is the case for another securitization measure, that is, the requirement of sufficient liquidity on the operators' side. A liquidity test can be applied to check if the experimenting company has enough assets to compensate in case of harm. Here we can borrow from the experience of the Australian Regulatory Sandbox where relevant Fintech firms are expected to provide evidence of sponsorship and a declaration that the firm 'has reasonable grounds to expect that it can operate its business'.⁵⁴ We again suggest that the state's financial credibility may be considered here as alternative, as an instrument to persuade the financially less affluent firms to join in the sandbox.

Potential damage to property—or more broadly to economic resources—can be translated into quantifiable risk relief measures fairly easily. That would be much harder in cases where the experiment may pose risks for life and limb or for political rights and freedoms. There the exit strategy has paramount importance, meaning that it should proactively be arranged that and planned how the experiment will be discontinued as soon as safeguarding the physical integrity of the person is otherwise not possible anymore. Ex post remedies shall be put in place as well, especially appropriate liability insurance in the way as pointed out above. The same would be advisable for potential harm to political rights and freedoms, although one may consider the situation less urgent. Here, an alternative scenario of the experiment can be prepared in case the risks materialise, so that the negative impact is reduced as much as possible.

We have already alluded to the issue of a waiver of rights by way of consent on the side of the individual in the capacity of consumer or otherwise impacted person. That would indeed be an important tool to secure that no impact on rights

⁵³ Council of Europe (CoE), Recommendation CM/Rec(2020)1 of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems. For a concrete tool, see Ministry of the Interior and Kingdom Relations (Netherlands) Impact Assessment Fundamental rights and algorithms (March 2022) on <www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms>.

⁵⁴ Australian Securities and Investments Commission, 'Further Measures to Facilitate Innovation in Financial Services' ASIC CP 260 (Consultation Paper No 260, June 2016).

goes without agreement from the individual holding the right. With that one can arrive at an acceptable risk reduction for consumers rights as well as political rights and privacy. It can be seen for instance in the Russian sandbox legislation, where participation in the experiment is subject to consent. This cannot work well for all types of experimentation, as we have shown. For that reason, experiments without explicit, free, and informed consent from all impacted persons should not take place in a consent-based sandbox practice. Implicit consent would in our view be unacceptable as risk-avoidance tactics. However, consent expressed through explicit and unambiguous behaviour, after sufficient information provision, could be found acceptable in specific cases.⁵⁵

As a final risk policy tool, we recommend the implementation of efficient mechanisms for quick and qualified resolution of the disputes arising from experimentation. Alternative dispute resolution, including online dispute resolution, could be arranged to provide all interested parties with a trustworthy and professional modality of conflict solution,⁵⁶ so as to avoid that the person suffering actual harm to fundamental rights has to go through lengthy court procedures at considerable cost before appropriate compensation is awarded.

6 Conclusion

In this chapter, we have tried to discover to what extent a compromise can be found between effective human rights protection on the one hand and experimentation in real life with AI-powered systems on the other hand. Attention was given mainly to the status quo in European States, both EU and non-EU. For the law of the ECHR the specifics of the relevant provisions' text and the case law of the ECtHR were studied. We can conclude now that a strict ban on experimentation cannot be based on this body of law. Nevertheless, the observance of a precautionary risk policy is conditional for a permissive attitude towards experimentation in regulatory sandboxes. We found good grounds for that compromise in the analysis of the positive obligations to protect fundamental rights and of the justification clauses in the ECHR, against the backdrop of the potential impact of allowing an experiment in regulatory sandboxes.

Our analysis of potential harm, in which we state that there is a spectrum of impact between no interference at all on one side and a full-blown violation of a right

⁵⁵ For examples not directly AI-related, see P Salvini and others, 'An Investigation on Legal Regulations for Robot Deployment in Urban Areas: A Focus on Italian Law' (2021) 24 Advanced Robotics 1901–17; and Yueh-Hsuan Weng and others, 'Intersection of "Tokku" Special Zone, Robots, and the Law: A Case Study on Legal Impacts to Humanoid Robots' (2015) 7 International Journal of Social Robotics 841–57.

⁵⁶ DB Ferreira and others, 'Arbitration Chambers and Trust in Technology Provider: Impacts of Trust in Technology Intermediated Dispute Resolution Proceedings' (2022) 68 Technology in Society 101872.

on the other, may be subject to discussion from the human rights theory angle. Can one express this in such varieties, or should we stick to the straightforward decision sequence (*viz.*: interference yes or no? If yes, justification yes or no? If no: violation.)? Related to that are the nuanced valuations of measures for relief and remedy that took the spectrum of impact as point of departure. This would indeed bring us to a deeper discussion on the theoretical core of our findings, while in this chapter we merely turned to the application on regulatory sandboxes for AI-powered systems.

Particularly for sandboxes in EU member states, more research is needed for a deeper understanding of the details of legally resilient practices. We have now taken the text of article 54 of the Draft AI Act as we find it in the proposal and have not in depth studied the text in juxtaposition to article 9 of the same Act. How these two provisions relate to the precautionary principle in the EU could be a good entry point for further research.

That research in the field of EU law could also focus more than we did on the potential breaches of data protection law when article 54 is applied to a concrete case. There is an indication that as such AI testing does not automatically cause GDPR infractions, but the question is not resolved to the full extent. Such research could also include the practice of testing algorithmic technologies that are intended to preserve or enhance privacy in the data processing going on in data-driven systems. One example of that would be advanced anonymisation techniques for health data. To what extent could this testing be acceptable without the standard consent of the data subject?

In section 3, we have very briefly mentioned the correlation between decent regulation of experimentation and trustworthiness of the tested AI system. There is enough indication in the literature for the assumption that clear and fair allocation of risk and responsibility contributes to trust.⁵⁷ However, an empirical comparative research into trust between *in vivo* tested and merely lab-tested AI applications to our knowledge is still lacking. The argument in this chapter nevertheless stands. The observance of the fundamental rights of impacted persons contributes to legally sound experimentation policies, capable of sustaining legal attacks. It would be a hallmark of the cautious introduction of AI in society.

⁵⁷ Philipp Schmidt, Felix Biessmann, and Timm Teubner, ‘Transparency and Trust in Artificial Intelligence Systems’ (2020) 29(4) *Journal of Decision Systems* 260–78; and Jeroen van de Hoven and others, ‘Towards a Digital Ecosystem of Trust: Ethical, Legal and Societal Implications’ (2021) *Opinio Juris in Comparatione* 131–56.

PART X
CONCLUSION

37

Conclusion

Alberto Quintavalla and Jeroen Temperman

The use of artificial intelligence (AI) has considerably affected most—if not all—domains of human life. Nowadays, it would be very challenging for an individual to avoid purportedly any interaction with AI applications. The goods we consume and the services we enjoy are inherently associated with the use of AI technology. It therefore cannot come as a surprise that AI technology may have an impact on human rights, too. This volume has attempted to comprehensively map and further analyse such an impact.

There are three main findings that stem from this volume. The first one—stating the obvious perhaps, but now at least documented in much detail—is that AI technology and human rights indeed have a double-edged relationship, whereby AI brings both significant risks and benefits to human rights protection. AI technology touches upon a broad range of issues related to several human rights and shows that its ramifications can be both beneficial and detrimental depending on the given human right considered. Hence, the deployment of AI technology does not establish a bilateral relationship with a single (specific set of) human right(s). The relationship between AI technology and human rights is a web of multilateral coexisting relationships.

The second—less trivial—main finding is that human rights norms can contribute to address this impact by acting as a prevention and mitigation system. This analysis in fact showed that despite the vast diversity among different AI applications, there are some common risks that require similar remedies at the technical and regulatory level. Human rights principles can provide an effective standard for measuring the societal acceptance of AI technology. Accordingly, the relationship between AI technology and human rights is not only unidirectional. Human rights can have an impact on AI, as well.

Both previous findings point at the fact that AI technology and human rights can be in principle both friends and foes. However, there is a third and more subtle main finding that complements our research inquiry: it is society who decides what type of impact AI technology makes—that is, becoming a friend or a foe of human rights. As so often in human relationships, this decision mostly depends on us. When meeting new individuals, would we like to turn them into friends or foes? It is usually us (say, society) to decide how to approach other humans. Humans per se are originally neutral, so is AI technology for human rights. This implies that

society has the task of reaching a clear vision on how to approach AI technology and, whenever needed, set certain boundaries via human rights norms.

These are the three main findings of the volume. Nonetheless, one may look for more specific answers on how we should deal with human rights in the AI context. The volume cannot currently offer detailed responses due to the novelty of the matter and the explorative nature of the research endeavour. The careful mapping exercise characterising this volume is already a considerable achievement in the nascent field of AI technology and human rights. In fact, there remain several challenges associated with the institutionalisation of this area.

Against this background, this volume can take an extra step. It points at a few observations underlying—what we believe to be—the three main challenges that will typify the AI and human rights scholarship in the following years. First, human rights norms predate AI technology. Despite starting in a similar period as discussed in the introduction, most of the human rights instruments were adopted before AI technology could unfold its full impact. Hence, the contemporary international human rights law still suffers from several gaps. One only needs to think of the limited regulation that the business sector receives and the consequent detrimental impact on human rights protection in the context of AI technology.

Second, the impact by AI technology is not only mixed but also pervasive in the human rights law (HRL) structure. It is very difficult to account for all the consequences that the development and deployment of a given AI application can have on human rights protection. This calls for a more careful analysis of the balancing exercise that courts will be required to conduct whenever they are presented with an alleged human right violation. In other words, the test of proportionality is likely to occupy an even greater role in the context of AI technology.

Third, the development and deployment of AI applications are very unevenly spread across the globe. Most of the industries developing AI technology tend to take input from marginalised communities and offer the output to wealthy segments of society. Some commentators have advanced the concept of 'AI colonialism' to describe that process. Hence, society should aim to reduce the glaring geographical inequalities within and between states.

To conclude, these are the three main challenges that are likely to characterise discussions at the crossroads of AI and human rights. Hopefully, in a few years, this volume can further incorporate some positive developments, integrate the analysis of the neglected human rights such as the right to education, and—finally—provide more certain answers. The ambition of this volume was limited to laying the groundwork for a mature discussion to be held among international and domestic institutions, business, and interested stakeholders. This research endeavour had become even more pressing given that a world without AI applications is harder to picture day after day.

Bibliography

All websites cited in the footnotes have last been accessed on 1 June 2022.

- Abbott R, *The Reasonable Robot: Artificial Intelligence and the Law* (CUP 2020).
- Abiteboul S and Dowek G, *The Age of Algorithms* (CUP 2020).
- Adams-Prassl J, 'What if Your Boss Was an Algorithm? Economic Incentives, Legal Challenges, and the Rise of Artificial Intelligence at Work' (2019) 41 Comparative Labor Law and Policy Journal 123.
- Adensamer A and Klausner L, 'Part Man, Part Machine, All Cop: Automation in Policing' (2021) 4 Frontiers in Artificial Intelligence 655486.
- Agre PE and Rotenberg M, 'Technology and Privacy: The New Landscape' (1997) 11 Harvard Journal of Law & Technology 3.
- Aizenberg E and Van den Hoven J, 'Designing for Human Rights in AI' (2020) 7(2) Big Data & Society 1.
- Ajunwa I, 'Algorithms at Work: Productivity Monitoring Applications and Wearable Technology as the New Data-Centric Research Agenda for Employment and Labor Law' (2018) 63 St Louis University Law Journal 21.
- , 'An Auditing Imperative for Automated Hiring Systems' (2021) 34 Harvard Journal of Law & Technology 622.
- , Crawford K, and Schultz J, 'Limitless Worker Surveillance' (2017) 105 California Law Review 735.
- Akter S, 'Algorithmic Bias in Data-Driven Innovation in the Age of AI' 60 (2021) International Journal of Information Management 102387.
- Albino V, Berardi U, and Dangelico RM, 'Smart Cities: Definitions, Dimensions, Performance, and Initiatives' (2015) 22(1) Journal of Urban Technology 3.
- Aletras N and others, 'Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective' (2016) 2 PeerJ Computer Science 1.
- Alhadef J, Van Alsenoy B, and Dumortier J, 'The Accountability Principle in Data Protection Regulation: Origin, Development and Future Directions' in D Guagnin and others (eds), *Managing Privacy through Accountability* (Palgrave Macmillan 2012) 49.
- Alikhademi K and others, 'A Review of Predictive Policing from the Perspective of Fairness' (2022) 30 Artificial Intelligence and Law 1.
- Allain J, 'The Jus Cogens Nature of Non-Refoulement' (2001) 13(4) International Journal of Refugee Law 533.
- Allen JA, 'The Color of Algorithms: An Analysis and Proposed Research Agenda for Deterring Algorithmic Redlining' (2019) 46 Fordham Urban Law Journal 219.
- Allhutter D and others, 'Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics are Made Effective' (2020) 3 Frontiers in Big Data 5.
- Alpaydin E, *Machine Learning* (MIT Press 2021).
- Amrute S, 'Of Techno-Ethics and Techno-Affects' (2019) 123 Feminist Review 56.
- Ananny M and Crawford K, 'Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 3 New Media & Society 973.

- Anderson JF, 'Big Brother or Little Brother-Surrendering Seizure Privacy for the Benefits of Communication Technology' (2012) 81 Mississippi Law Journal 895.
- Andrae ASG and Edler T, 'On Global Electricity Usage of Communication Technology: Trends to 2030' (2015) 6(1) Challenges 117.
- Aplin T, 'The Limits of EU Trade Secret Protection' in S Sandeen, C Rademacher, and A Ohly (eds), *Research Handbook on Information Law and Governance* (Edward Elgar 2021) 1.
- Arai Y, *The Margin of Appreciation Doctrine and the Principle of Proportionality in the Jurisprudence* (Intersentia 2002).
- Araujo AD and others, 'A Predictive Policing Application to Support Patrol Planning in Smart Cities', 2017 International Smart Cities Conference (ISC2) (IEEE 2017).
- Araujo T and others, 'In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence' (2020) 35 AI & Society 611.
- Are C, 'The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram' (2022) 22(8) Feminist Media Studies 2002.
- Aridor G, 'The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR' (2020) National Bureau of Economic Research 26900/2020.
- Aristova E and Grusic U, *Civil Remedies and Human Rights in Flux: Key Legal Developments in Selected Jurisdictions* (Bloomsbury 2022).
- Asaro P, 'What Should We Want from a Robot Ethic?' (2006) 6 International Review of Information Ethics 9.
- Ash J, Kitchin R, and Leszczynski A, 'Digital Turn, Digital Geographies?' (2018) 42(1) Progress in Human Geography 25.
- Askola H, 'Article 6: Right to Liberty and Security' in S Peers and others (eds), *The EU Charter of Fundamental Rights: A Commentary* (Hart 2021) 122.
- Aston V, 'State Surveillance of Protest and the Rights to Privacy and Freedom of Assembly: A Comparison of Judicial and Protester Perspectives' (2017) 8 European Journal of Law and Technology 1.
- Ateniese G and others, 'Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers' (2015) 10 International Journal of Security and Networks 137.
- Atkinson J, 'Workplace Monitoring and the Right to Private Life at Work' (2018) 81 Modern Law Review 688.
- Augenstein D, 'Negotiating the Hard/Soft Law Divide in Business and Human Rights: The Implementation of the UNGPs in the European Union' (2018) 9 Global Policy 254.
- Augustine of Hippo, *The Confessions* (RS Pine-Coffin tr, Penguin 1961).
- Baker RS and Hawn A, 'Algorithmic Bias in Education' (2022) 32 International Journal of Artificial Intelligence in Education 1052.
- Baldez L, 'The UN Convention to Eliminate All Forms of Discrimination Against Women (CEDAW): A New Way to Measure Women's Interests' (2011) 7 Politics & Gender 419.
- Balkin JM, 'Deconstructive Practice and Legal Theory' (1987) 96(4) Yale Law Journal 743.
- , 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 UC Davis Law Review 1149.
- , 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011.
- Bansak K and others, 'Improving Refugee Integration through Data-Driven Algorithmic Assignment' (2018) 359 Science 325.
- Bantekas I, Stein MA, and Anastasious D, *The UN Convention on the Rights of Persons with Disabilities: A Commentary* (OUP 2018).
- Bantema W and others, 'Black Box van Gemeentelijke Online Monitoring: Een Wankel Fundament onder een stevige praktijk' (Rijksuniversiteit Groningen 2021).

- Barak A, *Proportionality: Constitutional Rights and Their Limitations* (CUP 2012).
- Barcevičius E and others, 'Exploring Digital Government transformation in the EU—Analysis of the State of the Art and Review of Literature' (JRC Publications Office of the European Union 2019).
- Bar-Gill O, 'Algorithmic Price Discrimination: When Demand Is a Function of Both Preferences and (Mis) Perceptions' (2018) Harvard Public Law Working Paper No 18-32.
- Barnett I and others, 'Relapse Prediction in Schizophrenia through Digital Phenotyping: A Pilot Study' (2018) 43 *Neuropsychopharmacology* 1660.
- Barnett J and Treleaven P, 'Algorithmic Dispute Resolution—The Automation of Professional Dispute Resolution Using AI and Blockchain Technologies' (2018) 61(3) *The Computer Journal* 399.
- Barocas S and Selbst AD, 'Big Data's Disparate Impact' (2016) 104 *California Law Review* 671.
- Barry BM, *How Judges Judge: Empirical Insights Into Judicial Decision-Making* (Routledge 2020).
- Basu S and others, 'Legal Framework for Autonomous Agricultural Robots' (2020) 35 *AI & Society* 113.
- Bathaei Y, 'The Artificial Intelligence Black Box and the Failure of Intent and Causation' (2018) 31(2) *Harvard Journal of Law & Technology* 889.
- Bawden D and Robinson L, 'The Dark Side of Information: Overload, Anxiety and Other Paradoxes and Pathologies' (2009) 35 *Journal of Information Science* 180.
- Bekker S, 'Fundamental Rights in Digital Welfare States: The Case of SyRi in the Netherlands' (2019) 50 *Netherlands Yearbook of International Law* 289.
- Belenguer L, 'AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry' (2022) 2 *AI and Ethics* 771.
- Belkhir L and Elmeligi A, 'Assessing ICT Global Emissions Footprint: Trends to 2040 & Recommendations' (2018) 177 *Journal of Cleaner Production* 448.
- Bell W, *Foundation of Future Studies. History, Purposes, and Knowledge. Human Science for a New Era. Volume 1* (5th edn, Transaction 2009).
- Benedict TJ, 'The Computer Got It Wrong: Facial Recognition Technology and Establishing Probable Cause to Arrest' (2022) 79 (2) *Washington and Lee Law Review* 849.
- Benjamin SM, 'Algorithms and Speech' (2013) 161 *University of Pennsylvania Law Review* 1445.
- Bennett Moses L and Chan J, 'Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability' (2018) 28(7) *Policing and Society* 806.
- Bennett CJ, *Regulating Privacy* (Cornell UP 2018).
- Bennett CL and Keyes O, 'What Is the Point of Fairness? Disability, AI and the Complexity of Justice' (2020) 125 *ACM SIGACCESS Accessibility and Computing* 1.
- Benöhr I and Micklitz H-W, 'Consumer Protection and Human Rights' in G Howells, I Ramsay, and T Wilhelmsson (eds), *Handbook of Research on International Consumer Law* (2nd edn, Edward Elgar 2018) 16.
- Bergmann T and others, 'Developing a Diagnostic Algorithm for the Music-Based Scale for Autism Diagnostics (MUSAD) Assessing Adults with Intellectual Disability' (2019) 49 *Journal of Autism and Developmental Disorders* 3732.
- Berk R, 'An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism' (2017) 13 *Journal of Experimental Criminology* 193.
- Berle I, *Face Recognition Technology: Compulsory Visibility and Its Impact on Privacy and the Confidentiality of Personal Identifiable Images* (Springer 2020).

- Bertoni E and Kurre C, 'Surveillance and Privacy Protection in Latin America: Examples, Principles, and Suggestions' in Fred H Cate and James X Dempsey (eds), *Bulk Collection: Systemic Government Access to Private-Sector Data* (OUP 2017) 325.
- Besson S, 'Evolutions in Non-Discrimination Law within the ECHR and the ESC Systems: It Takes Two to Tango in the Council of Europe' (2021) 60 American Journal of Comparative Law 147.
- Bhandar B, *Colonial Lives of Property: Law, Land, and Racial Regimes of Ownership* (Duke UP 2018).
- Bigman YE and Gray K, 'People Are Averse to Machines Making Moral Decisions' (2018) 181 Cognition 21.
- Binns R, 'Data Protection Impact Assessments: A Meta-Regulatory Approach' (2017) 7 International Data Privacy Law 1.
- , 'Algorithmic Accountability and Public Reason' (2018) 31 Philosophy & Technology 543.
- Birch T, 'The Incarnation of Wilderness: Wilderness Areas Prisons' in M Oelschlaeger (ed), *Postmodern Environmental Ethics* (SUNY Press 1995) 39.
- Bisaso KR and others, 'A Survey of Machine Learning Applications in HIV Clinical Research and Care' (2017) 91 Computers in Biology and Medicine 366.
- Blank Y, 'Localism in the New Global Legal Order' (2007) 47 Harvard Journal of International Law 263.
- Blount K, 'Body Worn Cameras With Facial Recognition Technology: When It Constitutes a Search' (2015) 3 Criminal Law Practitioner 61.
- Bodo B and others, 'Tackling the Algorithmic Control Crisis: The Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents' (2017) 19 Yale Journal of Law & Technology 133.
- Boerman SC and others, 'Online Behavioral Advertising: A Literature Review and Research Agenda' (2017) 46 Journal of Advertising 363.
- Bolukbasi T and others, 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings' (2016) arXiv 1–3.
- Borgesius F, Helberger N, and Agustin R, 'The Perfect Match? A Closer Look at the Relationship Between EU Consumer Law and Data Protection Law' (2017) 54 Common Market Law Review 1427.
- Bostrom N, *Superintelligence: Paths, Dangers, Strategies* (OUP 2014).
- Boyd DR, 'Catalyst for Change: Evaluating Forty Years of Experience in Implementing the Right to a Healthy Environment' in JH Knox and R Pejan (eds), *The Human Right to a Healthy Environment* (CUP 2018) 17.
- Brantingham PJ, Valasik M, and Mohler G, 'Does Predictive Policing Lead to Biased Arrests? Results from a Randomized Controlled Trial' (2018) 5(1) Statistics and Public Policy 1.
- Broadhurst R, 'Developments in the Global Law Enforcement of Cyber-Crime' (2006) 29(3) Policing: An International Journal of Police Strategies & Management 408.
- Broomfield H and Reutter L, 'In Search of the Citizen in the Datafication of Public Administration' (2022) 9(1) Big Data & Society.
- Broude T, 'Behavioral International Law' (2014) 163 University of Pennsylvania Law Review 1099.
- Brown C, 'Investigating and Prosecuting Cybercrime: Forensic Dependencies and Barriers to Justice' (2015) 9(1) International Journal of Cyber Criminology 55.
- Brownsword R and Somsen H, 'Law, Innovation and Technology: Fast Forward to 2021' (2021) 13(1) Law, Innovation and Technology 1.

- Bruinsma FJ, 'Judicial Identities in the European Court of Human Rights' in A van Hoek and others (eds), *Multilevel Governance in Enforcement and Adjudication* (Intersentia 2006) 203.
- Bucher EL, Schou PK, and Waldkirch M, 'Pacifying the Algorithm: Anticipatory Compliance in the Face of Algorithmic Management in the Gig Economy' (2021) 28 Organization 44.
- Buçinca Z, Malaya M, and Gajos K, 'To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making' (2021) 5 Proceedings of the ACM on Human-Computer Interaction 1.
- Buolamwin J and Gebru T, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) 81 Proceedings of Machine Learning Research 77.
- Burk DL, 'Algorithmic Legal Metrics' (2021) 96(3) Notre Dame Law Review 1147.
- Burrell J, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society 1.
- Butenko A and Larouche P, 'Regulation for Innovativeness or Regulation of Innovation?' (2015) 7(1) Law, Innovation and Technology 52.
- Buvinic M and Levine R, 'Closing the Gender Data Gap' (2016) 13 Significance 34.
- Buyl M and others, 'Tackling Algorithmic Disability Discrimination in the Hiring Process: An Ethical, Legal and Technical Analysis' in 2022 ACM Conference on Fairness, Accountability, and Transparency (ACM 2022).
- Califano L and Fiorillo V, 'Videosorveglianza' in R Bifulco and others (eds), *Digesto-Discipline Pubblicistiche* (UTET 2015) 504.
- Calo R, 'Digital Market Manipulation' (2013) 82 George Washington Law Review 995.
- and others, 'Push, Pull, and Spill: A Transdisciplinary Case Study in Municipal Open Government' (2015) 30 Berkeley Technology Law Journal 1900.
- Calvi A, 'Gender, Data Protection & the Smart City: Exploring the Role of DPIA in Achieving Equality Goals' (2022) 19 European Journal of Spatial Development 24.
- Cannataci JA and Bonnici JPM, 'The End of the Purpose-Specification Principle in Data Protection?' (2010) 24 International Review of Law Computers and Technology 101.
- Canziani A, Paszke A, and Culurciello E, 'An Analysis of Deep Neural Network Models for Practical Applications' (2016) arXiv:1605.07678.
- Čapek K, *Rossum's Universal Robots* (D Wyllie tr, The Echo Library 2009).
- Caragliu A and others, 'Smart Cities in Europe' (2011) 18(2) Journal of Urban Technology 65.
- and Del Bo CF, 'Smart Cities and Urban Inequality' (2022) 56 Regional Studies 1097.
- Cardullo P and Kitchin R, 'Being a "Citizen" in the Smart City: Up and Down the Scaffold of Smart Citizen Participation in Dublin, Ireland' (2019) 84(1) GeoJournal 1.
- , 'Smart Urbanism and Smart Citizenship: The Neoliberal Logic of "Citizen-Focused" Smart cities in Europe' (2019) 37(5) Environment and Planning C: Politics and Space 813.
- Carlini N and others, 'The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks' in 28th USENIX Security Symposium (2019) 267.
- Carmien S and others, 'Socio-Technical Environments Supporting People with Cognitive Disabilities Using Public Transportation' (2005) 2 ACM Transactions on Computer-Human Interaction 30.
- Carolan E, 'Stars of Citizen CCTV: Video Surveillance and the Right to Privacy in Public Places' (2006) 13(1) Dublin University Law Journal 326.
- Carpenter M, 'Intellectual Property: A Human (Non-Corporate) Right' in D Keane and Y McDermott (eds), *The Challenge of Human Rights: Past, Present and Future* (Edward Elgar 2012) 322.

- Casey B, Farhangi A, and Vogl R, 'Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise' (2019) 34 Berkeley Technology Law Journal 170.
- Castagnola A, *Manipulating Courts in New Democracies: Forcing Judges off the Bench in Argentina* (Routledge 2018).
- Castelvecchi D, 'Is Facial Recognition Too Biased to Be Let Loose?' (2020) 587(7834) Nature 347.
- Cate FH and Mayer-Schönberger V, 'Notice and Consent in a World of Big Data' (2013) 3 International Data Privacy Law 67.
- , Cullen P, and Mayer-Schönberger V, *Data Protection Principles for the 21st Century: Revising the 1980 OECD Guidelines* (Microsoft 2013).
- Cepeda Espinosa MJ, 'Privacy' in M Rosenfeld and A Sajó (eds), *Oxford Handbook of Comparative Constitutional Law* (OUP 2012) 966.
- Cera R, Della Fina V, and Palmisano G, *The United Nations Convention on the Rights of Persons with Disabilities: A Commentary* (Springer 2017).
- Chan J, 'The Future of AI in Policing: Exploring the Sociotechnical Imaginaries' in JLM McDaniel and K Pease (eds), *Predictive Policing and Artificial Intelligence* (Routledge 2021) 41.
- Chandler D, *Ontopolitics in the Anthropocene: An Introduction to Mapping, Sensing and Hacking* (1st edn, Routledge 2018).
- Chazette L and Schneider K, 'Explainability as A Non-Functional Requirement: Challenges and Recommendations' (2020) 25(4) Requirements Engineering 493.
- Chen DL, 'Judicial Analytics and the Great Transformation of American Law' (2019) 27 Artificial Intelligence and Law 15.
- , 'Incremental AI' (2022) American Journal of Evaluation.
- Chen IY, Joshi S, and Ghassemi M, 'Treating Health Disparities with Artificial Intelligence' (2020) 26(1) Nature Medicine 16.
- Cheney-Lippold J, *We are Data Algorithms and the Making of Our Digital Selves* (NYU Press 2017).
- Chesterman S, 'Artificial Intelligence and the Limits of Legal Personality' (2020) 69 International and Comparative Law Quarterly 819.
- , *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law* (CUP 2021).
- Chiusi F and others (eds), *Automating Society* (Bertelsmann Stiftung and AlgorithmWatch 2020).
- Chivot E, 'The EU Needs to Reform the GDPR to Remain Competitive in the Algorithmic Economy' (Center for Data Innovation, 13 May 2019).
- Christensen J and others, 'Human Capital and Administrative Burden: The Role of Cognitive Resources in Citizen-State Interactions' (2020) 80 Public Administration Review 127.
- Christiaensen L and others, 'Viewpoint: The Future of Work in Agri-Food' (2021) 99 Food Policy 1.
- Christofi A and others, 'Data Protection, Control and Participation beyond Consent. "Seeking the Views" of Data Subjects in Data Protection Impact Assessments' in E Kosta, R Leenes, and I Kamara (eds), *Research Handbook on EU Data Protection Law* (Edward Elgar 2022) 503.
- Chu S-Y, *Do Metaphors Dream of Literal Sleep? A Science-Fictional Theory of Representation* (Harvard UP 2010).
- Churchland P, *Matter and Consciousness* (MIT Press 1999).

- Cima E, 'The Right to a Healthy Environment: Reconceptualizing Human Rights in the Face of Climate Change' (2022) 31(1) *Review of European, Comparative & International Environmental Law* 38.
- Citron DK, 'Sexual Privacy' (2019) 128 *Yale Law Journal* 1870.
- and Chesney R, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 *California Law Review* 1753.
- Clapham A, *Human Rights: A Very Short Introduction* (OUP 2007).
- Clarke G, 'The Evolving ASEAN Human Rights System: The ASEAN Human Rights Declaration of 2012' (2012) 11 *Northwestern Journal of Human Rights* 1.
- Clarke R, 'Privacy Impact Assessment: Its Origins and Developments' (2009) 25 *Computer Law & Security Review* 123.
- Cobbe J, 'Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decisionmaking' (2019) 39 *Legal Studies* 636.
- Coeckelbergh M, *Growing Moral Relations: Critique of Moral Status Ascription* (Palgrave 2012).
- Cofone I, 'Algorithmic Discrimination is an Information Problem' (2019) 70 *Hastings Law Journal* 1389.
- , 'Beyond Data Ownership' (2021) 43 *Cardozo Law Review* 501.
- and Strandburg K, 'Strategic Games and Algorithmic Secrecy' (2019) 64(4) *McGill Law Journal* 623.
- Coglianese C and Lai A, 'Algorithm vs. Algorithm' (2022) 71 *Duke Law Journal* 1281.
- and Lehr D, 'Regulating by Robot: Administrative Decision Making in the Machine-Learning Era' (2017) 105 *Georgetown Law Journal* 1147.
- Cohen IG and others, 'The European Artificial Intelligence Strategy: Implications and Challenges for Digital Health' (2020) 2(7) *The Lancet Digital Health* 367.
- Cohen J, 'What Privacy Is For' (2013) 126(7) *Harvard Law Review* 1904.
- Cole J, 'Empathy Needs a Face' (2001) 8(5–6) *Journal of Consciousness Studies* 51.
- Collins P, *Putting Human Rights to Work: Labour Law, the ECHR and the Employment Relation* (OUP 2022).
- and Marassi S, 'Is That Lawful? Data Privacy and Fitness Trackers in the Workplace' (2021) 37 *International Journal of Comparative Labour Law and Industrial Relations* 65.
- Colonna L, 'Data Mining and its Paradoxical Relationship to the Purpose of Limitation' in S Gutwirth, R Leenes, and P de Hert (eds), *Reloading Data Protection: Multidisciplinary Insights and Contemporary Challenges* (Springer 2014) 299.
- Connell R and Messerschmidt JW, 'Hegemonic Masculinity: Rethinking the Concept' (2005) 19(6) *Gender & Society* 829.
- Cook A and others, 'Towards Automatic Screening of Typical and Atypical Behaviors in Children with Autism' (2019) IEEE International Conference on Data Science and Advanced Analytics 504.
- Cook G and others, 'Clicking Clean: Who Is Winning the Race to Build a Green Internet' (2017) Greenpeace DC 5.
- Coulombel N and others, 'Substantial Rebound Effects in Urban Ridesharing: Simulating Travel Decisions in Paris, France' (2019) 71 *Transportation Research Part D: Transport and Environment* 110.
- Cowley R, Joss S, and Dayot Y, 'The Smart City and Its Publics: Insights from Across Six Cities' (2018) 11(1) *Urban Research and Practice* 53.
- Cox E, Royston S, and Selby J, 'The Impacts of Non-Energy Policies on the Energy System: A Scoping Paper' (2016) 100 UK Energy Research Centre 1.

- Crawford J and Brownlie I, *Brownlie's Principles of Public International Law* (OUP 2019).
- Crawford K, *Atlas of AI* (Yale UP 2022).
- Criado Perez C, *Invisible Women: Exposing Data Bias in a World Designed for Men* (Vintage 2019).
- Cusack S and Pusey L, 'CEDAW and the Rights to Non-Discrimination and Equality' (2013) 14 Melbourne Journal of International Law 54.
- Custers B and Ursic H, 'Workers' Privacy in a Digitized World under European Law' (2018) 39 Comparative Labor Law and Policy Journal 323.
- Daly E and May JR, 'Learning from Constitutional Environmental Rights' in JH Knox and R Pejan (eds), *The Human Right to a Healthy Environment* (CUP 2018) 42.
- Dameri RP, 'Smart City Definition, Goals and Performance' in R P Dameri (ed), *Smart City Implementation* (Springer 2017).
- Dann P, Riegner M, and Bönnemann M, 'The Southern Turn in Comparative Constitutional Law: An Introduction' in P Dann, M Riegner, and M Bönnemann (eds), *The Global South and Comparative Constitutional Law* (OUP 2020) 1.
- Dantcheva A, Elia P, and Ross A, 'What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics' (2015) IEEE Transactions on Information Forensics and Security 11.
- Daragh M and others, 'Effective Oversight of Large-Scale Surveillance Activities: A Human Rights Perspective' (2020) 11 Journal of National Security Law & Policy 749.
- Dash S and others, 'Big Data in Healthcare: Management, Analysis, and Future Prospects' (2019) 6 J Big Data 1.
- Datta A, 'The "Smart Safe City": Gendered Time, Speed, and Violence in the Margins of India's Urban Age' (2020) 110(5) Annals of the American Association of Geographers 1318.
- Daughety AF and Reinganum JF, 'Economic Analysis of Products Liability: Theory' in JH Arlen (ed), *Research Handbook on the Economics of Torts* (Edward Elgar 2013) 69.
- Davis MC, 'The Political Economy and Culture of Human Rights in East Asia' (2011) 1(1) Jindal Journal of International Affairs 48.
- De Alwis S and others, 'A Survey on Smart Farm Data, Application and Techniques' (2022) 138 Computers in Industry 103624.
- De Burca G, 'The EU in the Negotiation of the UN Disability Convention' (2010) 35(2) European Law Review 174.
- De Graaf M, Hindriks FA, and Hindriks KV, 'Who Wants to Grant Robots Rights?' (2021) Frontier in AI and Robotics 1.
- De Gregorio G, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (CUP 2022).
- and Dunn P, 'The European Risk-Based Approaches: Connecting Constitutional Dots in the Digital Age' (2022) 59(2) Common Market Law Review 473.
- and Stremlau N, 'Platform Governance at the Periphery: Moderation, Shutdowns and Intervention' in J Bayer and others (eds), *Perspectives on Platform Regulation. Concepts and Models of Social Media Governance Across the Globe* (Nomos 2021) 433.
- De Hert P, 'A Human Rights Perspective on Privacy and Data Protection Impact Assessments' in D Wright and P De Hert (eds), *Privacy Impact Assessment* (Springer 2012) 33.
- De la Feria R, 'Tax Fraud and Selective Law Enforcement' (2020) 47 Journal of Law and Society 266.
- De Laat PB, 'Algorithmic Decision-making Employing Profiling: Will Trade Secrecy Protection Render the Right to Explanation Toothless?' (2022) 24 Ethics and Information Technology 1.
- De Matos Pinto I, 'The Draft AI Act: A Success Story of Strengthening Parliament's Right of Legislative Initiative?' (2021) 22 ERA Forum 619.

- De Meeus C, ‘The Product Liability Directive at the Age of the Digital Industrial Revolution: Fit for Innovation?’ (2019) 8 *Journal of European Consumer and Market Law* 151.
- De Stefano V, ‘“Negotiating the Algorithm”: Automation, Artificial Intelligence, and Labor Protection’ (2019) 41 *Comparative Labor Law & Policy Journal* 15.
- De Vries A, ‘Bitcoin’s Growing Energy Problem’ (2018) 2(5) *Joule* 801.
- De Waal M and Dignum M, ‘The Citizen in the Smart City. How the Smart City Could Transform Citizenship’ (2017) 59(6) *Information Technology* 263.
- Deakin M and Al Waer H, ‘From Intelligent to Smart Cities’ (2011) 3(3) *Intelligent Buildings International* 140.
- Deibert RJ, ‘The Road to Digital Unfreedom: Three Painful Truths about Social Media’ (2019) 30(1) *Journal of Democracy* 25.
- Delfanti A, ‘Machinic Dispossession and Augmented Despotism: Digital Work in an Amazon Warehouse’ (2021) 23 *New Media & Society* 39.
- Dencik L and Kaun A, ‘Datafication and the welfare state’ (2020) 1(1) *Global Perspectives* 12912.
- Desai AN, ‘Artificial Intelligence: Promise, Pitfalls, and Perspective’ (2020) 323(24) *JAMA: Journal of the American Medical Association* 2448.
- Deutch S, ‘Are Consumer Rights Human Rights?’ (1994) 32(3) *Osgoode Hall Law Journal* 537.
- Dewey J, ‘The Historic Background of Corporate Legal Personality’ (1926) 35 *Yale Law Journal* 655.
- Dhar P, ‘The Carbon Impact of Artificial Intelligence’ (2020) 2(8) *Nature Machine Intelligence* 423.
- Oliva TD, Antonioli DM, and Gomes A, ‘Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online’ (2021) 25 *Sexuality & Culture* 700.
- Dietvorst BJ, Simmons JP, and Massey C, ‘Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err’ (2015) 114 *Journal of Experimental Psychology: General* 114.
- , ‘Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If they Can (even slightly) Modify Them’ (2018) 64 *Management Science* 1155.
- Dobbin F and Kalev A, ‘Multi-Disciplinary Responses to Susan Sturm’s The Architecture of Inclusion: Evidence from Corporate Diversity Programs’ (2007) 30 *Harvard Journal of Law and Gender* 279.
- Donahoe E and MacDuffee Metzger M, ‘Artificial Intelligence and Human Rights’ (2019) 30(2) *Journal of Democracy* 115.
- Došilović FK, Brčić M, and Hlupić N, ‘Explainable Artificial Intelligence: A Survey’ (2018) 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) 210.
- Dubber MD, Pasquale F, and Das S, *The Oxford Handbook of Ethics of AI* (OUP 2020).
- Duffield M, ‘The Resilience of the Ruins: Towards a Critique of Digital Humanitarianism’ (2016) 4 *Resilience* 147.
- , *Post-Humanitarianism: Governing Precarity in the Digital World* (John Wiley & Sons 2018).
- Durante M, *Ethics, Law and the Politics of Information: A Guide to the Philosophy of Luciano Floridi* (Springer 2017).
- Dworkin R, ‘Judicial Discretion’ (1963) 60 *The Journal of Philosophy* 624.
- , *Taking Rights Seriously* (Harvard UP 1981).

- Dzehtsiarou K and Schwartz A, 'Electing Team Strasbourg: Professional Diversity on the European Court of Human Rights and Why it Matters' (2020) 21 German Law Journal 621.
- Edel F, *The Length of Civil and Criminal Proceedings in the Case-Law of the European Court of Human Rights* (Human Rights Files No 16, Council of Europe 2007).
- Edwards L, 'Privacy, Security and Data Protection in Smart Cities: A Critical EU Law Perspective' (2016) 2(1) European Data Protection Law Review 28.
- and Veale M, 'Slave to the Algorithm? Why a "Right to an Explanation" is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law and Technology Review 77.
- Egbert S and Krasmann S, 'Predictive Policing: Not yet, but Soon Preemptive?' (2020) 30 Policing and Society 905.
- and Mann M, 'Discrimination in Predictive Policing: The (Dangerous) Myth of Impartiality and the Need for STS Analysis' in A Zavrnik and V Badalić (eds), *Automating Crime Prevention, Surveillance, and Military Operations* (Springer 2021) 2.
- Eisenberg M, 'The Conception that the Corporation is a Nexus of Contracts, and the Dual Nature of the Firm' (1998) 24 Journal of Corporation Law 819.
- Eneyew Ayalew Y, 'Untrodden Paths Towards the Right to Privacy in the Digital Era Under African Human Rights Law' (2022) 12(1) International Data Privacy Law 16.
- Engelbert J and Van Zoonen L, and H F, 'Excluding Citizens from the European Smart City: The Discourse Practices of Pursuing and Granting Smartness' (2019) 142 Technological Forecasting and Social Change 347.
- Englehart NA and Miller MK, 'The CEDAW Effect: International Law's Impact on Women's Rights' (2014) 13 Journal of Human Rights 22.
- Esposito R, *Persons and Things* (Zakiya Hanafi tr, Polity 2015).
- Estlund C, *Automation Anxiety: Why and How to Save Work* (OUP 2021).
- Eubanks V, *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor* (St Martin's Press 2018).
- Evans C, *Freedom of Religion under the European Convention on Human Rights* (OUP 2001).
- Fairfield JA, *Owned* (CUP 2017).
- Feldstein S, 'The Road to Digital Unfreedom: How Artificial Intelligence is Reshaping Repression' (2019) 30(1) Journal of Democracy 40.
- Ferguson AG, 'Predictive Policing and Reasonable Suspicion' (2012) 62 Emory Law Journal 259.
- , 'Facial Recognition and the Fourth Amendment' (2021) 105 Minnesota Law Review 1105.
- Ferré F, *Philosophy of Technology* (Prentice Hall 1998).
- Ferreira DB and others, 'Arbitration Chambers and Trust in Technology Provider: Impacts of Trust in Technology Intermediated Dispute Resolution Proceedings' (2022) 68 Technology in Society 101872.
- Ferreri M and Sanyal R, 'Digital Informalisation: Rental Housing, Platforms, and the Management of Risk' (2022) 37(6) Housing Studies 1035.
- Fields D and Rogers D, 'Towards a Critical Housing Studies Research Agenda on Platform Real Estate' (2021) 38 Housing, Theory and Society 72.
- Finck M and Biega A, 'Reviving Purpose Limitation and Data Minimisation in Personalisation, Profiling and Decision-Making Systems' (2021) Technology and Regulation 44.
- Fineman MA, 'The Vulnerable Subject: Anchoring Equality in the Human Condition' (2008) 20 Yale Journal of Law and Feminism 9.

- Firlej M and Taeihagh A, 'Regulating Human Control over Autonomous Systems' (2021) 15 *Regulation & Governance* 1071.
- Fjeld J and others, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI' (2020) Berkman Klein Center Research Publication 2020-1.
- Fletcher RR, Nakashimana A, and Olubeko O, 'Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health' (2021) 3 *Frontiers in Artificial Intelligence* 561802.
- Floridi L, *The Ethics of Information* (OUP 2013).
- , 'Robots, Jobs, Taxes, and Responsibilities' (2017) 30 *Philosophy & Technology* 4.
- , 'Artificial Intelligence, Deepfakes and a Future of Ectypes' (2018) 31 *Philosophy & Technology* 317.
- , 'Soft Ethics and the Governance of the Digital' (2018) 31(4) *Philosophy & Technology* 1.
- and Sanders, JW, 'On the Morality of Artificial Agents' (2004) 14 *Minds and Machine* 349.
- Fombu E, *Predictive Medicine: Artificial Intelligence and its Impact on Healthcare Business Strategy* (Business Expert Press 2020).
- Forgó N, Hänold S, and Schütze B, 'The Principle of Purpose Limitation and Big Data' in M Corrales, M Fenwick, and N Forgó (eds), *New Technology, Big Data and the Law* (Springer 2017) 34.
- Formosaa P and others, 'Medical AI and Human Dignity: Contrasting Perceptions of Human and Artificially Intelligent (AI) Decision Making in Diagnostic and Medical Resource Allocation Contexts' (2022) 133 *Computers in Human Behavior* 107296.
- Foss-Solbrekk K, 'Three Routes to Protecting AI Systems and their Algorithms under IP law: The Good, the Bad and the Ugly' (2021) 16(3) *Journal of Intellectual Property Law and Practice* 248.
- Foucault M, *Discipline and Punish: The Birth of the Prison* (2nd edn, Vintage 1995).
- Foulds JR and others, 'An Intersectional Definition of Fairness' in 2020 IEEE 36th International Conference on Data Engineering (ICDE) (IEEE 2020).
- Franks MA, '“Revenge Porn” Reform: A View from the Front Lines' (2017) 69 *Florida Law Review* 1251.
- , 'An-Aesthetic Theory: Adorno, Sexuality, and Memory' in RJ Heberle (ed), *Feminist Interpretations of Theodor Adorno* (Pennsylvania State UP 2006) 193.
- Frantzou E, 'The Horizontal Effect of the Charter: Towards an Understanding of Horizontality as a Structural Constitutional Principle' (2020) 22 *Cambridge Yearbook of European Legal Studies* 208.
- Frederickson G, *Social Equity and Public Administration: Origins, Developments, and Applications* (Routledge 2010).
- Fredman S, *Discrimination and Human Rights* (OUP 2004).
- Fredrikson M, Jha S, and Ristenpart T, 'Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures' in (2015) 22nd ACM SIGSAC Conference on Computer and Communications Security 1322.
- Free C and others, 'The Effectiveness of M-health Technologies for Improving Health and Health Services: A Systematic Review Protocol' (2010) 3 *BMC Research Notes* 1.
- Fried C, 'Privacy' (1968) 77 *Yale Law Journal* 475.
- Fry H, *Hello World: How to be Human in the Age of the Machine* (Random House 2018).
- Fudge J, 'Fragmenting Work and Fragmenting Organizations: The Contract of Employment and the Scope of Labour Regulation' (2006) 44 *Osgoode Hall Law Journal* 609.

- Fussey P and Murray D, 'Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology' (2019) *The Human Rights, Big Data and Technology Project*.
- , Davies B, and Innes M, 'Assisted Facial Recognition and the Reinvention of Suspicion and Discretion in Digital Policing' (2021) 61(2) *British Journal of Criminology* 325–44.
- Gabriel K, 'Feminist Revenge: Seeking Justice for Victims of Nonconsensual Pornography through "Revenge Porn" Reform' (2019) 44 *Vermont Law Review* 849.
- Galliot J, MacIntosh D, and Ohlin JD (eds), *Lethal Autonomous Weapons: Re-examining the Law and Ethics of Robotic Warfare* (OUP 2021).
- Gambs S and others, 'Privacy and Ethics—Understanding the Convergences and Tensions for the Responsible Development of Machine Learning' (Office of the Privacy Commissioner of Canada 2021).
- Gantchev V, 'Data Protection in the Age of Welfare Conditionality: Respect for Basic Rights or a Race to the Bottom?' (2019) 21(1) *European Journal of Social Security* 3.
- Garapon A and Lassègue J, *Justice Digitale* (PUF 2018).
- García Segura LA, 'European Cybersecurity: Future Challenges from a Human Rights Perspective' in J Martín Ramírez and J Biziewski (eds), *Security and Defence in Europe* (Springer 2020) 35.
- Gates KA, *Our Biometric Future* (NYU Press 2011).
- Gautier A, Ittoo A, and Van Cleynenbreugel P, 'AI Algorithms, Price Discrimination and Collusion: A Technological, Economic and Legal Perspective' (2020) 50 *European Journal of Law and Economics* 405.
- Geiger C, 'Implementing Intellectual Property Provisions in Human Rights Instruments: Towards a New Social Contract for the Protection of Intangibles' in C Geiger (ed), *Research Handbook on Human Rights and Intellectual Property* (Edward Elgar 2015) 661.
- , 'Reconceptualizing the Constitutional Dimension of Intellectual Property' in P Torremans (ed), *Intellectual Property and Human Rights* (Wolters Kluwer 2015) 115.
- Gellers JC, *Rights for Robots: Artificial Intelligence, Animal and Environmental Law* (Routledge 2021).
- and Gunkel D, 'Artificial Intelligence and International Human Rights Law: Implications for Humans and Technology in the 21st Century and Beyond' in A Zwitter and O Gstrein (eds), *Handbook on the Politics and Governance of Big Data and Artificial Intelligence* (Edward Elgar 2023).
- Gellert R, 'Understanding the Notion of Risk in the General Data Protection Regulation' (2018) 34 *Computer Law & Security Review* 279.
- Geraci R, 'Apocalyptic AI: Religion and the Promise of Artificial Intelligence' (2008) 76 *Journal of the American Academy of Religion* 138.
- Gerards J, *General Principles of the European Convention on Human Rights* (CUP 2019).
- , 'The Fundamental Rights Challenges of Algorithms' (2019) 37(3) *Netherlands Quarterly of Human Rights* 205.
- and Xenidis R, *Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Anti-Discrimination Law* (European Commission 2020).
- and Zuiderveen Borgesius F, 'Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence' (2022) 20 *Colorado Technology Law Journal* 1.
- Gerdes CJ and Thornton SM, 'Implementable Ethics for Autonomous Vehicles' in M Maurer and others (eds), *Autonomes Fahren* (Springer 2015) 87.

- Gerke S, Minssen T, and Cohen G, 'Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare' in A Bohr and K Memarzadeh (eds), *Artificial Intelligence in Healthcare* (Academic Press 2020) 295.
- and others, 'The Need for a System View to Regulate Artificial Intelligence/Machine Learning-Based Software as Medical Device' (2020) 3(1) Digital Medicine 1.
- Ghassemi M and others, 'Practical Guidance on Artificial Intelligence for Health-Care Data' (2019) 1(4) *The Lancet Digital Health* 157.
- Gibson J, 'Where Have You Been? CGI Film Stars and Reanimation Horrors' (2020) 10(1) Queen Mary Journal of Intellectual Property 1.
- Giegerich T, *The European Union as Protector and Promoter of Equality* (Springer 2020).
- Giffinger R and others, *Smart Cities: Ranking of European Medium-Sized Cities* (Vienna University of Technology 2007).
- Gilbert A and Thomas A, *The Amazonian Era: How Algorithmic Systems Are Eroding Good Work* (Institute for the Future of Work 2021).
- Gillespie T, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale UP 2018).
- Ginsburg T, 'Authoritarian International Law?' (2020) 114(2) American Journal of International Law 221.
- Girasa R, *Artificial Intelligence as a Disruptive Technology* (Palgrave Macmillan 2020).
- Giuffrida N and others, 'Optimization and Machine Learning Applied to Last-Mile Logistics: A Review' (2022) 14(9) Sustainability 5329.
- Goggin G, 'Disability and Digital Inequalities: Rethinking Digital Divides with Disability Theory' in M Ragnedda and G Muschert (eds), *Theorizing Digital Divides* (Routledge 2018) 63.
- , Ellis K, and Hawkins W, 'Disability at the Centre of Digital Inclusion: Assessing a New Moment in Technology and Rights' (2019) 5(3) Communication Research and Practice 290.
- Goldenein J, *Monitoring Laws: Profiling and Identity in the World State* (CUP 2019).
- Gomes Pereira V, 'Using Supervised Machine Learning and Sentiment Analysis Techniques to Predict Homophobia in Portuguese Tweets' (PhD Dissertation, Fundação Getulio Vargas, 2018).
- Gonzalez MT, 'Habeas Data: Comparative Constitutional Interventions from Latin America Against Neoliberal States of Insecurity and Surveillance' (2015) 90(2) Chicago-Kent Law Review 641.
- Gonzalez-Salzberg D and Hodson L (eds), *Research Methods for International Human Rights Law. Beyond the Traditional Paradigm* (Routledge 2020).
- Goodman M, *Future Crimes: A Journey to the Dark Side of Technology and How to Survive It* (Transworld 2015).
- and others, 'Size and Distribution of Transgender and Gender Nonconforming Populations: A Narrative Review' (2019) 48(2) Endocrinology and Metabolism Clinics of North America 303.
- Gordon J-S, 'Artificial Moral and Legal Personhood' (2021) 26 AI & Society 470.
- Gorwa R, Binns R, and Katzenbach C, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 1.
- Gosse CE and Burkell J, 'Politics and Porn: How News Media Characterizes Problems Presented by Deepfakes' (2020) 37(5) Critical Studies in Media Communication 497.
- Goswami M, 'Algorithms and Freedom of Expression' in W Barfield (ed), *The Cambridge Handbook of the Law of Algorithms* (CUP 2020) 558.

- Grant MJ and Booth A, 'A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies' (2009) 26(2) *Health Information & Libraries Journal* 91.
- Gravett WH, 'Digital Neocolonialism: The Chinese Surveillance State in Africa' (2022) 30(1) *African Journal of International and Comparative Law* 39.
- Gray K, Young L, and Waytz A, 'Mind Perception Is the Essence of Morality' (2012) 23(2) *Psychological Inquiry* 101.
- Grear A, 'Challenging Corporate Humanity Legal Disembodiment, Embodiment and Human Rights' (2007) 7 *Human Rights Law Review* 511.
- Green M, 'Speech Acts' in *The Stanford Encyclopedia of Philosophy* (2021).
- Greenfield A, *Everyware: The Dawning Age of Ubiquitous Computing* (New Riders 2006).
- Greenleaf G, *Asian Data Privacy Laws: Trade and Human Rights Perspectives* (OUP 2014).
- and Cottier B, 'International and Regional Commitments in African Data Privacy Laws: A Comparative Analysis' (2022) 44 *Computer Law & Security Review* 105638.
- Grimmelmann J, 'The Virtues of Moderation' (2015) 17 *Yale Journal of Law and Technology* 42.
- Grochowski M and others, 'Algorithmic Price Discrimination and Consumer Protection: A Digital Arms Race?' (2022) *Technology and Regulation* 36.
- Gromova E and Ivanc T, 'Regulatory Sandboxes (Experimental Legal Regimes for digital innovations) for BRICS' (2020) 7(2) *BRICS Law Journal* 10–36.
- Gross R, 'Information Property Rights and The Information Commons' in R Frank Jørgensen, EJ Wilson, and WJ Drake (eds), *Human Rights in the Global Information Society* (MIT Press 2006) 109.
- Grosse Ruse-Khan H, 'Assessing the Need for a General Public Interest Exception in the TRIPS Agreement' in A Kur (ed), *Intellectual Property in a Fair World Trade System* (Edward Elgar 2011) 167.
- Grother P, Ngan ML, and Hanaoka K, *Ongoing Face Recognition Vendor Test (FRYT) Part 2: Identification* (National Institute of Standards and Technology 2018).
- , Ngan ML, and Hanaoka K, *Face Recognition Vendor Test (FYRT) Part 3: Demographic Effects* (National Institute of Standards and Technology 2019).
- and others, *Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification* (National Institute of Standards and Technology 2022).
- Grozdanovski L, 'In Search of Effectiveness and Fairness in Proving Algorithmic Discrimination in EU Law' (2021) 58(1) *Common Market Law Review* 99.
- Gunkel DJ, *Robot Rights* (MIT Press 2018).
- Guo C, Bond CA, and Narayanan A, *The Adoption of New Smart-Grid Technologies: Incentives, Outcomes, and Opportunities* (Rand Corporation 2015).
- Gupta M and others, 'Security and Privacy in Smart Farming: Challenges and Opportunities' (2020) 8 *IEEE Access* 34569.
- Gwagwa A and others, 'Artificial Intelligence (AI) Deployments in Africa: Benefits, Challenges and Policy Dimensions' (2020) 26 *The African Journal of Information and Communication* 1.
- Haber E, 'Racial Recognition' (2021) 43(1) *Cardozo Law Review* 71.
- Hacker P, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law' (2018) 55(4) *Common Market Law Review* 1143.
- and others, 'Explainable AI Under Contract and Tort Law: Legal Incentives and Technical Challenges' (2020) 28 *Artificial Intelligence and Law* 415.

- Haggerty K and Ericson R, 'The Surveillant Assemblage: Surveillance, Crime and Social Control' (2000) 51(4) *British Journal of Sociology* 605.
- Hamilton M, 'Investigating Algorithmic Risk and Race' (2021) 5 *UCLA Criminal Justice Law Review* 530.
- , 'Predictive Policing through Risk Assessment' in JLM McDaniel and K Pease (eds), *Predictive Policing and Artificial Intelligence* (Routledge 2021) 60.
- Hämmig O, 'Health Risks Associated with Social Isolation in General and in Young, Middle and Old Age' (2019) 14 *PLoS One* e0219663.
- Hamon R and others, 'Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making' (2022) 17 *IEEE Computational Intelligence Magazine* 72.
- Hansen J and others, *Assessment of the EU Member States' Rules on Health Data in the Light of GDPR* (Publications Office of the EU 2021).
- Harbo T-I, *The Function of Proportionality Analysis in European Law* (Brill/Nijhoff 2015).
- Harpur P, *Discrimination, Copyright and Equality* (CUP 2019).
- Harris DJ and others, *Harris, O'Boyle & Warbrick: Law of the European Convention on Human Rights* (3rd edn, OUP 2014).
- Harrison K, 'Who Is the Assumed User in the Smart City?' in V Angelakis and others (eds), *Designing, Developing, and Facilitating Smart Cities: Urban Design to IoT Solutions* (Springer 2017) 17.
- Hart HLA, *The Concept of Law* (2nd edn, OUP 1997).
- Hart O, 'An Economist's Perspective on the Theory of the Firm' (1989) 89 *Columbia Law Review* 1757.
- Hartzog W, 'What is Privacy? That is the Wrong Question' (2021) 88 *University of Chicago Law Review* 1677.
- Hassel H and Cedergren A, 'Integrating Risk Assessment and Business Impact Assessment in the Public Crisis Management Sector' (2021) 56 *International Journal of Disaster Risk Reduction* 102136.
- Hayward P and Rahn A, 'Opening Pandora's Box: Pleasure, Consent and Consequence in the Production and Circulation of Celebrity Sex Videos' (2015) 2(1) *Porn Studies* 1.
- Heine K, 'Interjurisdictional Competition and the Allocation of Constitutional Rights: A Research Note' (2006) 26 *International Review of Law and Economics* 33.
- Heine K and Quintavalla A, 'Bridging the accountability gap of artificial intelligence—what can be learned from Roman law?' (2023) *Legal Studies* 1.
- Heinrichs B, 'Discrimination in the Age of Artificial Intelligence' (2022) 37(1) *AI & Society* 143.
- Heins V, 'Human Rights, Intellectual Property and Struggles for Recognition' (2007) 9 *Human Rights Review* 213.
- Helberger N, 'Profiling and Targeting Consumers in the Internet of Things—A New Challenge for Consumer Law' in R Schulze and D Staudenmayer (eds), *Digital Revolution: Challenges for Contract Law in Practice* (Nomos 2016) 135.
- and others, 'Implications of AI-Driven Tools in the Media for Freedom of Expression' (Council of Europe 2020).
- Helper L and Austin G, *Human Rights and Intellectual Property: Mapping the Global Interface* (CUP 2011).
- Helveston MN, 'Consumer Protection in the Age of Big Data' (2015) 93 *Washington University Law Review* 859.

- Henry N and Powell A, 'Embodied Harms: Gender, Shame, and Technology-Facilitated Sexual Violence' (2015) 21(6) *Violence Against Women* 758.
- , Flynn A, and Powell A, 'Policing Image-based Sexual Abuse: Stakeholder Perspectives' (2018) 19(6) *Police Practice and Research* 565.
- and Witt A, 'Governing Image-Based Sexual Abuse: Digital Platform Policies, Tools, and Practices' in J Bailey and others (eds), *The Emerald International Handbook of Technology-Facilitated Violence and Abuse* (Emerald 2021) 749.
- Hermalin BE and Katz ML, 'Privacy, Property Rights and Efficiency: The Economics of Privacy as Secrecy' (2006) 4 *Quantitative Marketing and Economics* 209.
- Herrero JG and others, *Study on Face Identification Technology for its Implementation in the Schengen Information System* (Publications Office of the European Union 2019).
- Hevelke A and Nida-Rümelin J, 'Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis' (2015) 21 *Science and Engineering Ethics* 619.
- Heyns C, 'Human Rights and the Use of Autonomous Weapons Systems (AWS) During Domestic Law Enforcement' (2016) 38 *Human Rights Quarterly* 350.
- Hildebrandt M, *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology* (Edward Elgar 2015).
- , 'Algorithmic Regulation and the Rule of Law' (2018) 376 *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 20170355.
- , 'Smart Technologies' (2020) 9(4) *Internet Policy Review* 1.
- Hildt E, 'Artificial Intelligence: Does Consciousness Matter?' (2019) 10 *Frontiers in Psychology* 1.
- Hirsch E, 'Kripke's Argument Against Materialism' in RC Koons and G Bealer (eds), *The Waning of Materialism* (OUP 2010) 115.
- Hirschl R, *City, State: Constitutionalism and the Megacity* (OUP 2020).
- and Schachar A, 'Spatial Statism' (2019) 18(1) *ICON* 387.
- Hiutunen E, 'Good Sources of Weak Signals: A Global Study of Where Futurists Look For Weak Signals' (2008) 12 *Journal of Future Studies* 21.
- Hobbes T, *Leviathan* (JCA Gaskin ed, OUP 2008).
- Hoffman F, 'Facing South: On the Significance of An/Other Modernity in Comparative Constitutional Law' in P Dann, M Rieger, and M Bönnemann (eds), *The Global South and Comparative Constitutional Law* (OUP 2020) 41.
- Hoffmann AL, 'Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse' (2019) 22 *Information, Communication & Society* 900.
- Hohfeld W, *Fundamental Legal Conceptions as Applied in Judicial Reasoning* (Yale UP 1920).
- Hohmann J, *The Right to Housing: Law, Concepts, Possibilities* (Hart 2013).
- , 'Conceptualising Domestic Servitude as a Violation of the Human Right to Housing and Reframing Australian Policy Responses' (2022) 31(1) *Griffith Law Review* 98.
- Hollands RG, 'Will the Real Smart City Stand Up?' (2008) 12(3) *City* 302.
- Hoofnagle CJ, Van der Sloot B, and Zuiderveen Borgesius F, 'The European Union General Data Protection Regulation: What It Is and What It Means' (2018) 28(1) *Information & Communications Technology Law* 65.
- Hornung G and Schnabel C, 'Data Protection in Germany I: The Population Census Decision and the Right to Informational Self-Determination' (2009) 25 *Computer Law & Security Review* 84.
- Hosny A and Aerts HJWL, 'Artificial Intelligence for Global Health' (2019) 366(6468) *Science* 955.
- Houwing L, 'Opinions: Stop the Creep of Biometric Surveillance Technology' (2020) 6 *European Data Protection Law Review* 174.

- Howells G, Twigg-Flesner C, and Wilhelmsson T, *Rethinking EU Consumer Law* (Taylor & Francis 2017).
- Hubbard DW, *How to Measure Anything. Finding the Value of 'Intangibles' in Business* (Wiley 2014).
- Huckvale K, Svetha V, and Christensen H, 'Toward Clinical Digital Phenotyping: A Timely Opportunity to Consider Purpose, Quality, and Safety' (2019) 2(1) NPJ Digital Medicine 1.
- Hudson R, 'Ethical Investing: Ethical Investors and Managers' 2005 (15) Business Ethics Quarterly 641.
- Humble K, 'International Law, Surveillance and the Protection of Privacy' (2021) 25(1) International Journal of Human Rights 1.
- Hume D, *Enquiries Concerning Human Understanding and Concerning the Principles of Morals* (3rd edn, OUP 1975).
- Hylton KN, 'The Law and Economics of Products Liability' (2012) 88 Notre Dame Law Review 2457.
- Iglesias M and others, *Intellectual Property and Artificial Intelligence: A Literature Review* (2019) (Publications Office of the European Union 2021).
- Iglesias M, Schamulia S, and Anderberg A, 'Artificial Intelligence and Intellectual Property—A Literature Review' (Publication Office of the European Union 2021).
- Introna L and Wood D, 'Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems' (2004) 2 Surveillance & Society 177.
- Isetti G and others (eds), *Religion in the Age of Digitalization: From New Media to Spiritual Machines* (Routledge 2021).
- Jabłonowska A and others, 'Consumer Law and Artificial Intelligence: Challenges to the EU Consumer Law and Policy Stemming from the Business' Use of Artificial Intelligence—Final report of the ARTSY project' (2018) EUI Department of Law Research Paper 2018/11.
- Jacques L, 'Facial Recognition Technology and Privacy: Race and Gender—How to Ensure the Right to Privacy is Protected' (2021) 23 San Diego International Law Journal 111.
- Jain A, Bolle R, and Pankanti S, 'Introduction to Biometrics' in A Jain, R Bolle, and S Pankanti (eds), *Biometrics* (Springer 1996) 1.
- Jak N and Bastiaans S, 'De Betekenis van de AVG voor Geautomatiseerde Besluitvorming door de Overheid: Een Black Box voor een Black Box?' (2018) 40 Nederlands Juristenblad 3018.
- Jansen F, Sánchez-Monedero J, and Dencik L, 'Biometric Identity Systems in Law Enforcement and the Politics of (Voice) Recognition: The Case of SiiP' (2021) 8(2) Big Data & Society 4.
- Janssen H, 'An Approach for a Fundamental Rights Impact Assessment to Automated Decision-Making' (2020) 10 International Data Privacy Law 76.
- Jasserand C, 'Avoiding Terminological Confusion Between the Notions of "Biometrics" and "Biometric Data": An Investigation into the Meanings of the Terms From a European Data Protection and a Scientific Perspective' (2016) 6(1) International Data Privacy Law 63.
- Jauhar A, 'Facing up to the Risks of Automated Facial-Recognition Technologies in Indian Law Enforcement' (2020) 16 Indian Journal of Law and Technology 1.
- Jayawickrama N, *The Judicial Application of Human Rights Law: National, Regional and International Jurisprudence* (CUP 2017).
- Jennings Saul C and Gebauer H, 'Digital Transformation as an Enabler for Advanced Services in the Sanitation Sector' (2018) 10 Sustainability 752.

- Jewell M, 'Contesting the Decision: Living in (and Living with) the Smart City' (2018) 32(2) International Review of Law, Computers & Technology 210.
- Johnson AM and Axinn S, 'The Morality of Autonomous Robots' (2013) 12(2) Journal of Military Ethics 129.
- Johnson DG and Miller KW, 'Un-Making Artificial Moral Agents' (2008) 10 Ethics and Information Technology 123.
- Johnson PA, 'Models of Direct Editing of Government Spatial Data: Challenges and Constraints to the Acceptance of Contributed Data' (2017) 44(2) Cartography and Geographic Information Science 128.
- Jones C, 'Law Enforcement Use of Facial Recognition: Bias, Disparate Impacts on People of Color, and the Need for Federal Legislation' (2020) 22 North Carolina Journal of Law & Technology 777.
- Jones L, 'A Philosophical Analysis of AI and Racism' (2020) 13(1) Stance 36.
- Jones N, 'The Information Factories' (2018) 561(7722) Nature 163.
- Jørgensen RF, 'The Right to Express Oneself and to Seek Information' in RF Jørgensen, EJ Wilson, and WJ Drake (eds), *Human Rights in the Global Information Society* (MIT Press 2006) 55.
- , 'Data and Rights in the Digital Welfare State: The Case of Denmark' (2021) Information, Communication & Society 1.
- and Pedersen AM, 'Online Service Providers as Human Rights Arbiters' in M Taddeo and L Floridi (eds), *The Responsibilities of Online Service Providers* (Springer 2017) 31.
- Joseph S and Castan M, 'Freedom from Arbitrary Detention: Article 9' in S Joseph and M Castan (eds), *The International Covenant on Civil and Political Rights: Cases, Materials, and Commentary* (3rd edn, OUP 2013) 341.
- Joss S and others, 'The Smart City as Global Discourse: Storylines and Critical Junctures across 27 Cities' (2019) 26(1) Journal of Urban Technology 3.
- Kahneman D, Sibony O, and Sunstein CR, *Noise: A Flaw in Human Judgment* (Little Brown 2021).
- Kaminski M, 'The Right to Explanation, Explained' (2019) 4 Berkeley Technology Law Journal 1.
- and Malgieri G, 'Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations' (2021) 11 International Data Privacy Law 20.
- and Witnov S, 'The Conforming Effect: First Amendment Implications of Surveillance, Beyond Chilling Speech' (2015) 49 University of Richmond Law Review 483.
- Kant I, *Groundwork for the Metaphysic of Morals* (M Gregor and J Timmermann trs, 2nd edn, CUP 2012).
- Kapczynski A, 'The Law of Informational Capitalism' (2020) 129(5) Yale Law Journal 1460.
- Karanasiou AP and Pinotsis D, 'Towards a Legal Definition of Machine Intelligence: The Argument for Artificial Personhood in the Age of Deep Learning' 2017 Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law 119.
- Karnouskos S, 'Symbiosis with Artificial Intelligence via the Prism of Law, Robots, and Society' (2021) 30(1) Artificial Intelligence and Law 1.
- Karp DJ, 'What Is the Responsibility to Respect Human Rights? Reconsidering the "Respect, Protect, and Fulfill" Framework' (2020) 12(1) International Theory 83.
- Katyal SK, 'Private Accountability in the Age of Artificial Intelligence' (2018) 66 UCLA Law Review 115.
- Kaufman H, 'Fear of Bureaucracy: A Raging Pandemic' (1981) 41(1) Public Administration Review 1.

- Kaun A, 'Suing the Algorithm: The Mundanization of Automated Decision-Making in Public Services through Litigation' (2021) 25(14) *Information, Communication & Society* 2046.
- Kay RS, 'The State Action Doctrine, the Public-Private Distinction, and the Independence of Constitutional Law Symposium on the State Action Doctrine' (1993) 10 *Constitutional Commentary* 329.
- Kearns M and Roth A, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (OUP 2020).
- Keats Citron D, 'Technological Due Process' (2008) 85(6) *Washington University Law Review* 1249.
- and Pasquale F, 'The Scored Society: Due Process for Automated Predictions' (2014) 89 *Washington Law Review* 1.
- Keenan S, 'From Historical Chains to Derivative Futures: Title Registries as Time Machines' (2019) 20 *Social & Cultural Geography* 283.
- Keizer A-G, Tiemeijer W, and Bovens M, *Why Knowing What to Do is Not Enough: A Realistic Perspective on Self-Reliance* (WRR 2019).
- Kelly L, *Surviving Sexual Violence* (1st edn, University of Minnesota Press 1988).
- Kelly-Lyth A, 'Challenging Biased Hiring Algorithms' (2021) 41 *Oxford Journal of Legal Studies* 889.
- , 'Dispatch 39—European Union: The AI Act and Algorithmic Management' (2021) *Comparative Labor Law & Policy Journal* 1.
- Keramitsoglou I, Cartalis C, and Kiranoudis CT, 'Automatic Identification of Oil Spills on Satellite Images' (2006) 21(5) *Environmental Modelling & Software* 640.
- Keyes O, 'The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition' (2018) 2 *Proceedings of the ACM on Human-Computer Interaction* 88.
- Khaliq U and Churchill R, 'The European Committee of Social Rights: Putting Flesh on the Bare Bones of the European Social Charter' in M Langford (ed), *Social Rights Jurisprudence: Emerging Trends in International and Comparative Law* (CUP 2008) 429.
- Khatua A and others, 'Tweeting in Support of LGBT?: A Deep Learning Approach', CoDS-COMAD 2019: *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data* (January 2019).
- Kickbusch I and others, 'The Lancet and Financial Times Commission on Governing Health Futures 2030: Growing Up in a Digital World' (2021) 398b(10312) *The Lancet* 1727.
- Kim PT, 'Data-Driven Discrimination at Work' (2017) 58 *William & Mary Law Review* 857.
- Kinchin N, 'Technology, Displaced? The Risks and Potential of Artificial Intelligence for Fair, Effective, and Efficient Refugee Status Determination' (2021) 37(3) *Law in Context* 5.
- Kindt E, *Privacy and Data Protection Issues of Biometric Applications* (Springer 2016).
- , 'Having Yes, Using No? About the New Legal Regime for Biometric Data' (2018) 34(3) *Computer Law & Security Review* 1.
- Kirwan CG and Zhiyong F, *Smart Cities and Artificial Intelligence* (Elsevier 2020).
- Kiskens T, *Towards an International Political Economy of Artificial Intelligence* (Springer Nature 2022).
- Kissinger HA, Schmidt E, and Huttenlocher D, *The Age of AI And Our Human Future* (John Murray 2021).
- Kitchin R and Dodge M, 'The (In)Security of Smart Cities: Vulnerabilities, Risks, Mitigation, and Prevention' (2019) 26(2) *Journal of Urban Technology* 47.
- Kleinberg J and others, 'Human Decisions and Machine Predictions' (2018) 133 *The Quarterly Journal of Economics* 273.

- Klonick K, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2017) 131 Harvard Law Review 1598.
- Kloza D and others, 'The Concept of Impact Assessment' in JP Burgess and D Kloza (eds), *Border Control and New Technologies. Addressing Integrated Impact Assessment* (Uitgeverij ASP 2021) 31.
- Knox JH, 'Horizontal Human Rights Law' (2008) 102 American Journal of International Law 1.
- Koch C, 'Proust among the Machines' (2019) 321 Scientific American 46.
- Kocsis E, 'Deepfakes, Shallowfakes, and the Need for a Private Right of Action' (2022) 126(2) Dickinson Law Review 621.
- Kohli R and others, 'Artificial Intelligence Technology to Help Students with Disabilities: Promises and Implications for Teaching and Learning' in A Singh and others (eds), *Handbook of Research on Critical Issues in Special Education for School Rehabilitation Practices* (IGI Global 2021) 238.
- Kohn NA, 'Vulnerability Theory and the Role of Government' (2014) 26(1) Yale Journal of Law and Feminism 1.
- Kołacz MK, Quintavalla A, and Yalnazov O, 'Who Should Regulate Disruptive Technology?' (2019) 10(1) European Journal of Risk Regulation 4.
- Königs P, 'Government Surveillance, Privacy, and Legitimacy' (2022) 35(8) Philosophy and Technology 8.
- Koning ME, 'The Purpose And Limitations Of Purpose Limitation' (DPhil Thesis, Radboud University Nijmegen 2020).
- Koops B-J, 'The Concept of Function Creep' (2021) 13(1) Law, Innovation and Technology 29.
- Kosinski M, 'Facial Recognition Technology Can Expose Political Orientation from Naturalistic Facial Images' (2021) 11(1) Scientific Reports 100.
- Kosta E, 'Algorithmic State Surveillance: Challenging the Notion of Agency in Human Rights' (2020) 16 Regulation & Governance 212.
- Kostka G, Steinacker L, and Meckel M, 'Between Security and Convenience: Facial Recognition Technology in the Eyes of Citizens in China, Germany, the United Kingdom, and the United States' (2021) 30(6) Public Understanding of Science 671.
- Kostopoulos L, 'The Emerging Artificial Intelligence Wellness Landscape: Benefits and Potential Areas of Ethical Concern' (2018) 55 (1) California Western Law Review 235.
- Koulu R, 'Blockchains and Online Dispute Resolution: Smart Contracts as an Alternative to Enforcement' (2016) 13 SCRIPTed 40.
- , 'Crafting Digital Transparency: Implementing Legal Values into Algorithmic Design' (2021) 8 Critical Analysis of Law 81.
- Kouroupis K, 'Facial Recognition: A Challenge for Europe or a Threat to Human Rights?' (2021) 2021 European Journal of Privacy Law & Technologies 142.
- Kramer M, Simmonds NE, and Steiner H, *A Debate Over Rights: Philosophical Enquiries* (OUP 1998).
- and others, 'When Do People Want AI to Make Decisions?' (2018) Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 204.
- Krotoszynski Jr RJ, *Privacy Revisited: A Global Perspective on the Right to Be Left Alone* (OUP 2016).
- Krüger M, 'The Impact of Video Tracking Routines on Crowd Behaviour and Crowd Policing' in L Melgaço and J Monaghan (eds), *Protests in the Information Age: Social Movements, Digital Practices and Surveillance* (Routledge 2018) 135.

- Krupiy T and McLeod Rogers J, 'Mapping Artificial Intelligence and Human Intersections; Why We Need New Perspectives on Harm and Governance in Human Rights' in A O'Donoghue, R Houghton, and S Wheatle (eds), *Research Handbook on Global Governance* (Edward Elgar 2023).
- Kuner C and others, 'Expanding the Artificial Intelligence-Data Protection Debate' (2018) 8 International Data Privacy Law 290.
- Kwok R, 'AI Empowers Conservation Biology' (2019) 567(7746) Nature 133.
- La Fors-Owczynik K, 'Profiling "Anomalies" and the Anomalies of Profiling: Digitalized Risk Assessments of Dutch Youth and the New European Data Protection Regime' in S Adams, N Purtova, and R Leenes (eds), *Under Observation: The Interplay Between eHealth and Surveillance* (Springer 2017) 107.
- Lacoste A and others, 'Quantifying the Carbon Emissions of Machine Learning' (2019) arXiv:1910.09700.
- Lageson SE, McElrath S, and Palmer KE, 'Gendered Public Support for Criminalizing "Revenge Porn"' (2019) 14(5) Feminist Criminology 560.
- LaGrandeur K, 'How Safe Is Our Reliance on AI, and Should We Regulate It?' (2021) 1 AI and Ethics 93.
- Laidlaw E, 'Myth or Promise? The Corporate Social Responsibilities of Online Service Providers for Human Rights' in M Taddeo and L Floridi (eds), *The Responsibilities of Online Service Providers* (Springer 2017) 135.
- LaMonaga J, 'A Break from Reality: Modernizing Authentication Standards for Digital Video Evidence in the Era of Deepfakes' (2020) 69(6) American University Washington College of Law 1945.
- Langford M, Behn D, and Lie R, 'Computational Stylometry: Predicting the Authorship of Investment Arbitration Awards' in R Whalen (ed), *Computational Legal Studies: The Promise and Challenge of Data-Driven Research* (Edward Elgar 2020) 53.
- Lanzing M, 'The Transparent Self' (2016) 18 Ethics and Information Technology 9.
- , 'Strongly Recommended: Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies' (2019) 32 Philosophy & Technology 549.
- Larsson KK, 'Digitization or Equality: When Government Automation Covers Some, But Not All Citizens' (2021) 38(1) Government Information Quarterly 101547.
- Larsson S and Heintz F, 'Transparency in Artificial Intelligence' (2020) 9 Internet Policy Review.
- Latonero M, 'Governing Artificial Intelligence: Upholding Human Rights & Dignity' (2018) Data & Society 1.
- and Agarwal A, 'Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar' (Carr Center for Human Rights Policy Harvard Kennedy School, Harvard University 2021).
- Lazar J and Stein MA, *Disability, Human Rights, and Information Technology* (University of Pennsylvania Press 2017).
- Lazaro C and Le Métayer D, 'Control over Personal Data: True Remedy or Fairy Tale?' (2015) 12 SCRIPTed 3.
- Lazcoz G and De Hert P, 'Humans in the GDPR and AIA Governance of Automated and Algorithmic Systems. Essential Pre-Requisites Against Abdicating Responsibilities' (2022) 8 Brussels Privacy Hub Working Paper.
- Leavy K, 'Chilling Effects and Unequal Subjects: A Response to Jonathon Penney's Understanding Chilling Effects' (2022) 106 Minnesota Law Review 395.
- Lee K-F, *AI 2041—Ten Visions for the Future* (Currency 2021).

- Lee MSA and others, 'Defining the Scope of AI ADM System Risk Assessment' in E Kosta and I Kamara (eds), *Research Handbook on EU Data Protection Law* (Edward Elgar 2022) 405.
- Lehr D and Ohm P, 'Playing with the Data: What Legal Scholars Should Learn about Machine Learning' (2017) 51 UC Davis Law Review 653.
- Lenaerts K and others, *Digital Platform Work and Occupational Safety and Health: A Review* (European Agency for Safety and Health at Work 2021).
- Leslie D and others, *Artificial Intelligence, Human Rights, Democracy, and The Rule of Law: A Primer* (Council of Europe and The Alan Turing Institute 2021) 38.
- Lester A, 'Universality versus Subsidiarity: A Reply' (1998) 3 European Human Rights Law Review 73.
- Letaifa SB, 'How to Strategize Smart Cities: Revealing the SMART Model' (2015) 68(7) Journal of Business Research 1414.
- Li D and others, 'Evaluating the Energy Efficiency of Deep Convolutional Neural Networks on cpus and gpus' (2016) IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloudSocialCom-SustainCom) 477.
- Lichter A and others, 'The Long-Term Costs of Government Surveillance: Insights from Stasi Spying in East Germany' (2021) 19(2) Journal of the European Economic Association 741.
- Liebowitz J (ed), *Data Analytics and AI* (Routledge 2020).
- Lima G and others, 'Collecting the Public Perception of AI and Robot Rights' (2020) 4 Proceedings of the ACM on Human-Computer Interaction 4.
- Limon M, 'The Politics of Human Rights, the Environment, and Climate Change at the Human Rights Council' in JH Knox and R Pejan (eds), *The Human Right to a Healthy Environment* (CUP 2018) 189.
- Lippi M and others, 'CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service' (2019) 27 Artificial Intelligence and Law 118.
- and others, 'Consumer Protection Requires Artificial Intelligence' (2019) 1 Nature Machine Intelligence 168.
- Liu H-Y, 'The Power Structure of Artificial Intelligence' (2018) 10 Law, Innovation and Technology 197.
- Livingston S and Risso M, 'The Future Impact of Artificial Intelligence on Humans and Human Rights' (2019) 33(2) Ethics & International Affairs 141.
- Lock T and Martin D, 'Article 47 CFR: Right to an Effective Remedy and to a Fair Trial' in M Kellerbauer, M Klamert, and J Tomkin (eds), *The EU Treaties and the Charter of Fundamental Rights: A Commentary* (OUP 2019) 2214.
- Lohsse S, Schulze R, and Staudenmayer D, 'Liability for Artificial Intelligence' in S Lohsse, R Schulze, and D Staudenmayer (eds), *Liability for Artificial Intelligence and the Internet of Things* (Hart 2019) 11.
- Londjani P and others, 'AIA: AI and EU Agriculture' (EC Science and Knowledge Service Joint Research Centre 2020).
- Longoni C, Bonezzi A, and Morewedge CK, 'Resistance to Medical Artificial Intelligence' (2019) 4 Journal of Consumer Research 629.
- Louridas P, *Algorithms* (MIT Press 2020).
- Luo Y and Guo R, 'Facial Recognition in China: Current Status, Comparative Approach and the Road Ahead' (2021) 25 Journal of Law and Social Change 153.
- Lurie G, 'Proportionality and the Right to Equality' (2020) 21 (2) German Law Journal 174.

- Lütz F, 'Gender Equality and Artificial Intelligence in Europe. Addressing Direct and Indirect Impacts of Algorithms on Gender-Based Discrimination' (2022) 23(1) ERA Forum 37.
- Lyon D, 'Surveillance as Social Sorting: Computer Codes and Mobile Bodies' in D Lyon (ed), *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination* (1st edn, Routledge 2002) 13.
- Macdonald SJ and others, ‘“The Invisible Enemy”: Disability, Loneliness and Isolation’ (2018) 33 Disability & Society 1138.
- Macfarlane K, ‘Disability without Documentation’ (2021) 90 Fordham Law Review 60.
- Maddocks S, ‘“A Deepfake Porn Plot Intended to Silence Me”: Exploring Continuities Between Pornographic and “Political” Deep Fakes’ (2020) 7 Porn Studies 415.
- Maddox TM, Rumsfeld JS, and Payne PRO, ‘Questions for Artificial Intelligence in Health Care’ (2019) 321(1) JAMA: Journal of the American Medical Association 31.
- Madsen MR and Spano R, ‘Authority and Legitimacy of the European Court of Human Rights: Interview with Robert Spano, President of the European Court of Human Rights’ (2020) 1(2) European Convention on Human Rights Law Review 165.
- Main EC and Walker TG, ‘Choice Shifts and Extreme Behavior: Judicial Review in the Federal Courts’ (1973) 91 The Journal of Social Psychology 215.
- Makulilo AB, ‘Myth and Reality of Harmonisation of Data Privacy Policies in Africa’ (2015) 31(1) Computer Law and Security Review 78.
- , ‘The Future of Data Protection in Africa’ in AB Makulilo (ed), *African Data Privacy Laws* (Springer 2016) 371.
- Malgieri G, ‘Trade Secrets v Personal Data: A Possible Solution for Balancing Rights’ (2016) 6 International Data Privacy Law 102.
- , ‘The Concept of Fairness in the GDPR’ (Conference on Fairness, Accountability, and Transparency, 2020) 156.
- and Comandé G, ‘Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation’ (2017) 7 International Data Privacy Law 243.
- and Niklas J, ‘Vulnerable Data Subjects’ (2020) 37 Computer Law & Security Review 105415.
- Mallan A, *Lekenbescherming in het Bestuursprocesrecht* (Wolf Legal 2014).
- Manheim K and Kaplan L, ‘Artificial Intelligence: Risks to Privacy and Democracy’ (2019) 21 Yale Journal of Law & Technology 106.
- Mania K, ‘The Legal Implications and Remedies Concerning Revenge Porn and Fake Porn: A Common Law Perspective’ (2020) 24 Sexuality & Culture 2079.
- Mann M and Smith M, ‘Automated Facial Recognition Technology: Recent Developments and Approaches to Oversight’ (2017) 40 University of New South Wales Law Journal 121.
- Mantelero A, ‘AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment’ (2018) 34 Computer Law & Security Review 754.
- , *Report on Artificial Intelligence Artificial Intelligence and Data Protection: Challenges and Possible Remedies* (Council of Europe 2019).
- , *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI* (Springer 2022).
- and Esposito MS, ‘An Evidence-Based Methodology for Human Rights Impact Assessment (HRIA) in the Development of AI Data-Intensive Systems’ (2021) 41 Computer Law & Security Review 105561.
- Mantouvalou V, ‘“I Lost My Job over a Facebook Post: Was That Fair?” Discipline and Dismissal for Social Media Activity’ (2019) 35 International Journal of Comparative Labour Law and Industrial Relations 101.

- Mantziari D, 'Sadistic Scopophilia in Contemporary Rape Culture: I Spit On Your Grave and the Practice of "Media Rape"' (2018) 18(3) Feminist Media Studies 397.
- Manuel EJ and Wachter RM, 'Artificial Intelligence in Health Care' (2019) 321(23) JAMA: Journal of the American Medical Association 2281.
- Marcus JL and others, 'Use of Electronic Health Record Data and Machine Learning to Identify Candidates for HIV Pre-Exposure Prophylaxis: A Modelling Study' (2019) 10(6) The Lancet HIV 688.
- Marelli L, Lievrouw E, and Van Hoyweghen I, 'Fit for Purpose? The GDPR and the Governance of European Digital Health' (2020) 41 Policy Studies 453.
- Mark R, 'Ethics of Public Use of AI and Big Data: The Case of Amsterdam's Crowdedness Project' (2019) 2(2) The ORBIT Journal 1.
- Martin JA and Fargo AL, 'Anonymity as a Legal Right: Where and Why It Matters' (2015) 16 North Carolina Journal of Law & Technology 311.
- Martínez VC and Castillo GP, 'Historia del Fake Audiovisual: Deepfake y la Mujer en un Imaginario Falsificado y Perverso' (2019) 24(2) Historia y Comunicación Social 505.
- Marx G, 'Coming to Terms: The Kaleidoscope of Privacy and Surveillance' in B Roessler and D Mokrosinska (eds), *Social Dimensions of Privacy: Interdisciplinary Perspectives* (CUP 2015) 36.
- Marx J and Tiefensee C, 'Of Animals, Robots and Men' (2015) 40 Historical Social Research 70.
- Mateescu A and Nguyen A, 'Algorithmic Management in the Workplace' (2019) Data & Society 1.
- Matwyshyn AM, 'The Internet of Bodies' (2019) 61 William & Mary Law Review 77.
- Mayer C, 'The Future of the Corporation: Towards Humane Business' (2018) 6 Journal of the British Academy 1.
- Mayer-Schonberger V and Cukier K, *Big Data: A Revolution That Will Transform How We Live, Work and Think* (John Murray 2013).
- Mayor A, *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology* (Princeton UP 2018).
- Mayson SG, 'Bias In, Bias Out' (2019) 128 Yale Law Journal 2218.
- McCorquodale R and Tse M, 'Artificial Intelligence Impacts: A Business and Human Rights Approach' (2021) 26 Communications Law 11.
- McElroy E, 'Property as Technology: Temporal Entanglements of Race, Space, and Displacement' (2020) 24 City 112.
- McGill J and Salyzyn A, 'Judging by Numbers: How Will Judicial Analytics Impact the Justice System and Its Stakeholders?' (2021) 44 The Dalhousie Law Journal 249.
- McGlynn C and Rackley E, 'Image-Based Sexual Abuse' (2017) 37(3) Oxford Journal of Legal Studies 534.
- McGregor L, Murray D, and Ng V, 'International Human Rights Law as a Framework for Algorithmic Accountability' (2019) 68 International & Comparative Law Quarterly 309.
- McKenna E, Richardson I, and Thomson M, 'Smart Meter Data: Balancing Consumer Privacy Concerns with Legitimate Applications' (2012) 41 Energy Policy 807.
- McNally P and Inayatullah S, 'The Rights of Robots: Technology, Culture and Law in the 21st Century' (1988) 20 Futures 123.
- Medvedeva M, Vols M, and Wieling M, 'Using Machine Learning to Predict Decisions of the European Court of Human Rights' (2020) 28 Artificial Intelligence and Law 237.
- Méret F and Alston P, *The United Nations and Human Rights: A Critical Appraisal* (OUP 2020).

- Mehrabi N and others, 'A Survey on Bias and Fairness in Machine Learning' (2021) 54(6) ACM Computing Surveys 1.
- Meijer A and Wessels M, 'Predictive Policing: Review of Benefits and Drawbacks' (2019) 42(12) International Journal of Public Administration 1031.
- Melgaço L and Van Brakel R, 'Smart Cities as Surveillance Theatre' (2021) 19(2) Surveillance & Society 244.
- Meron T, 'The Meaning and Reach of the International Convention on the Elimination of all Forms of Racial Discrimination' (1985) 79 American Journal of International Law 283.
- Meskys E and others, 'Regulating Deep Fakes: Legal and Ethical Considerations' (2020) 15(1) Journal of Intellectual Property Law & Practice 24.
- Micklitz H-W, Reich N, and Rott P, *Understanding EU Consumer Law* (Intersentia 2009) 21–26.
- , Pałka P, and Panagis Y, 'The Empire Strikes Back: Digital Control of Unfair Terms of Online Services' (2017) 40 Journal of Consumer Policy 367.
- Milano S, Taddeo M, and Floridi L, 'Recommender Systems and Their Ethical Challenges' (2020) 35 AI & Society 957.
- Minssen T and others, 'Regulatory Responses to Medical Machine Learning' (2020) 7(1) Journal of Law and the Biosciences 1.
- Mittelstadt BD and others, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3(2) Big Data and Society 1.
- Möller K, 'Proportionality: Challenging the Critics' (2012) 10(3) International Journal of Constitutional Law 711.
- Molnar P, 'Technology on the Margins: AI and Global Migration Management from a Human Rights Perspective' (2019) 305 Cambridge International Law Journal 306.
- , 'Robots and Refugees: the Human Rights Impacts of Artificial Intelligence and Automated Decision-Making in Migration' in M McAuliffe (ed), *Research Handbook on International Migration and Digital Technology* (Edward Elgar 2021) 136.
- and Gill L, *Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System* (University of Toronto Press 2018).
- Monfort GF, *Nuevas Perspectivas del Poder de Dirección y Control del Empleador* (Editorial Bomarzo 2016).
- Moore GE, 'Cramming More Components onto Integrated Circuits' (1965) 38 Electronics 114.
- Moore P, *OSH and the Future of Work: Benefits and Risks of Artificial Intelligence Tools in Workplaces* (European Agency for Safety and Health at Work 2019).
- Moreira J, Carvalho A, and Horvath T, *A General Introduction to Data Analytics* (Wiley 2019).
- Moss G and Ford H, 'How Accountable Are Digital Platforms?' in WH Dutton (ed), *A Research Agenda for Digital Politics* (Edward Elgar 2020) 97.
- Müller J and Kerényi A, 'The Need for Trust and Ethics in the Digital Age: Sunshine and Shadows in the FinTech World' (2019) 3 Financial and Economic Review 5.
- Murdoch B, 'Privacy and Artificial Intelligence: Challenges for Protecting Health Information in a New Era' (2021) 22(5) BMC Medical Ethics 122.
- Murovana T, 'Media and Information Literacy & Artificial Intelligence' in I Kushchuk and T Demirel (eds), *Artificial Intelligence: Media and Information Literacy, Human Rights and Freedom of Expression* (UNESCO IIITE, TheNextMinds 2020).
- Musila G, 'The Right to an Effective Remedy Under the African Charter on Human and Peoples' Rights' (2006) 6 African Human Rights Law Journal 442.

- Nagel T, 'What Is It Like to Be a Bat?' (1974) 83 *The Philosophical Review* 435.
- Nakamura K, 'My Algorithms Have Determined You're Not Human: AI-ML, Reverse Turing Tests, and the Disability Experience' in 21st International ACM SIGACCESS Conference on Computers and Accessibility (2019).
- Narayanan A, 'AI Snake Oil, Pseudoscience and Hype' in F Kaltheuner (ed), *Fake AI* (Meatspace Press 2021) 24.
- Naruniec J and others, 'High-Resolution Neural Face Swapping for Visual Effects' (2020) 39(4) *Computer Graphics Forum* 173.
- Nemitz P, 'Constitutional Democracy and Technology in the Age of Artificial Intelligence' (2018) 376 *Philosophical Transactions of the Royal Society* 11.
- Neroni Rezende I, 'Facial Recognition in Police Hands: Assessing the "Clearview Case" from a European Perspective' (2020) 11 *New Journal of European Criminal Law* 375.
- Newton A, *The Business of Human Rights: Best Practice and the UN Guiding Principles* (Routledge 2019).
- Nicola H and Powell A, 'Technology-Facilitated Sexual Violence: A Literature Review of Empirical Research' (2018) 19(2) *Trauma, Violence & Abuse* 195.
- Nissenbaum H, 'Protecting Privacy in an Information Age: The Problem of Privacy in Public' (1998) 17(5/6) *Law and Philosophy* 559.
- Noble SU, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York UP 2018).
- Nolan J, 'The Corporate Responsibility to Respect Rights: Soft Law or Not Law?' in S Deva and D Bilchitz (eds), *Human Rights Obligations of Business: Beyond the Corporate Responsibility to Respect* (CUP 2013) 138.
- Noorbakhsh-Sabet N and others, 'Artificial Intelligence Transforms the Future of Health Care' (2019) 132(7) *American Journal of Medicine* 795.
- Nordling L, 'A Fairer Way Forward for AI in Health Care' (2019) 573(7775) *Nature* S103.
- Norouzzadeh MS and others, 'Automatically Identifying, Counting, and Describing Wild Animals in Camera-Trap Images with Deep Learning' (2018) 115(25) *Proceedings of the National Academy of Sciences* E5716.
- Noto La Diega G, 'Against the Dehumanisation of Decision-Making: Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information' (2018) 9 *Journal of Intellectual Property, Information Technology and E-Commerce Law* 3.
- Nouwens M and others, 'Dark Patterns After the GDPR: Scraping Consent Pop-Ups and Demonstrating their Influence' (2020) *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.
- Novelli C, Bongiovanni G, and Sartor G, 'A Conceptual Framework for Legal Personality and its Application to AI' (2021) 13(2) *Jurisprudence* 194.
- Nti EK and others, 'Environmental Sustainability Technologies in Biodiversity, Energy, Transportation and Water Management using Artificial Intelligence: A Systematic Review' (2022) *Sustainable Futures* 100068.
- Ntoutsis E and others, 'Bias in Data-Driven Artificial Intelligence Systems: An Introductory Survey' (2020) 10 *WIREs Data Mining and Knowledge Discovery* 1356.
- Nuijens J, 'Bijstandsfraude Voorspellen Met Big Data' (2017) 1 *Sociaal Web*.
- O'Connor J and Shaw K, 'What Next for the Creative City' (2014) 5(3) *Culture and Society* 165.
- O'Mellin L, 'Software and Shovels: How the Intellectual Property Revolution is Undermining Traditional Concepts of Property' (2007) 76 *University of Cincinnati Law Review* 143.

- O’Neil C, *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy* (Broadway Books 2017).
- Obermeyer Z and others, ‘Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations’ (2019) 366(6464) Science 447.
- Öderström O, Paasche T, and Klauser F, ‘Smart Cities as Corporate Storytelling’ (2014) 18(3) City 307.
- Ohlhausen M and Okuliar A, ‘Competition, Consumer Protection, and the Right (Approach) to Privacy’ (2015) 80 Antitrust Law Journal 121.
- Ohm P, ‘Changing the Rules: General Principles for Data Use and Analysis’ in J Lane and others (eds), *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (CUP 2014) 100.
- Onnela J-P, ‘Opportunities and Challenges in the Collection and Analysis of Digital Phenotyping Data’ (2022) 46(1) Neuropsychopharmacology 45.
- Oomen B, ‘Human Rights Cities: The Politics of Bringing Human Rights Home to the Local Level’ in J Handmaker and K Arts (eds), *Mobilising International Law for ‘Global Justice’* (CUP 2020) 208.
- Oren M and Alterman R, ‘The Right to Adequate Housing Around the Globe: Analysis and Evaluation of National Constitutions’ in S Agarwal (ed), *Rights and the City: Problems, Progress and Practice* (University of Alberta Press 2022).
- Ostergaard R, ‘Intellectual Property: A Universal Human Right?’ (1999) 21 Human Rights Quarterly 156.
- Oswald M and others, ‘Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and “Experimental” Proportionality’ (2018) 27(2) Information & Communications Technology Law 223.
- Owens D, *Shaping the Normative Landscape* (2nd edn, OUP 2014).
- Packin NG, ‘Disability Discrimination Using Artificial Intelligence Systems and Social Scoring: Can We Disable Digital Bias?’ (2021) 8 Journal of International and Comparative Law 487.
- Pagallo U, ‘Killers, Fridges, and Slaves: A Legal Journey in Robotics’ (2011) 26 AI & Society 347.
- Panch T and others, ‘Artificial Intelligence, Machine Learning and Health Systems’ (2018) 8(2) Journal of Global Health 1.
- and others, ‘Artificial Intelligence: Opportunities and Risks for Public Health’ (2019) 1(1) Lancet Digital Health 13, 13.
- Parasuraman R and Manzey D, ‘Complacency and Bias in Human Use of Automation: An Attentional Integration’ (2010) 52(3) Human Factors 381.
- Parikh RB, Teeple S, and Navathe AS, ‘Addressing Bias in Artificial Intelligence in Health Care’ (2019) 322(24) JAMA: Journal of the American Medical Association 377.
- Pariser E, *The Filter Bubble: What the Internet is Hiding From You* (Penguin 2011).
- Parsons C, ‘The (in)effectiveness of Voluntarily Produced Transparency Reports’ (2019) 58 Business & Society 103.
- Pasquale F, *The Black Box Society: The Secret Algorithms that Control Money and Information* (Harvard UP 2015).
- , *New Laws of Robotics: Defending Human Expertise in the Age of AI* (Harvard UP 2020).
- Patelet F and others, ‘Combating Depression in Students Using an Intelligent ChatBot: A Cognitive Behavioral Therapy’ in 2019 IEEE 16th India Council International Conference (INDICON) (IEEE 2019).

- Paterson JM and Maker Y, 'AI in the Home: Artificial Intelligence and Consumer Protection Law' in E Lim and P Morgan (eds), *The Cambridge Handbook of Private Law and Artificial Intelligence* (CUP 2022) 2.
- Paton Simpson E, 'Privacy and the Reasonable Paranoid: The Protection of Privacy in Public Places' (2000) 50(3) University of Toronto Law Journal 305.
- Pavis M, 'Rebalancing Our Regulatory Response to Deepfakes with Performers' Rights' (2021) 27(4) *Convergence: The International Journal of Research into New Media Technologies* 974.
- Pavone V, Santiago Gomez E, and Jaquet-Chifelle D-O, 'A Systemic Approach to Security: Beyond the Tradeoff Between Security and Liberty' (2016) 12(4) *Democracy and Security* 225.
- Pearl J and Mackenzie D, *The Book of Why: The New Science of Cause and Effect* (Penguin 2018).
- Pechenik Gieseke A, '"The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography' (2020) 73(5) *Vanderbilt Law Review* 1479.
- Peng X, 'Coping with Population Ageing in Mainland China' (2021) 17 *Asian Population Studies* 1.
- Penney J, 'Understanding Chilling Effects' (2022) 106 *Minnesota Law Review* 1451.
- Perez CC, *Invisible Women. Exposing Data Bias in a World Designed for Men* (Vintage 2020).
- Peters J, 'The "Sovereigns of Cyberspace" and State Action: The First Amendment's Application—Or Lack Thereof—to Third-Party Platforms' (2017) 32 *Berkeley Technology Law Journal* 989.
- Peters MA and Besley T, 'Critical Philosophy of the Postdigital' (2019) 1 *Postdigital Science and Education* 29.
- Peukert C and others, 'Regulatory Spillovers and Data Governance: Evidence from the GDPR' (2022) 41 *Marketing Science* 318.
- Pfleeger CP and Pfleeger SL, *Analyzing Computer Security: A Threat/Vulnerability/Countermeasure Approach* (Prentice Hall Professional 2012).
- Phan HD, 'The Evolution Towards an ASEAN Human Rights Body' (2008) 9 *Asia-Pacific Journal on Human Rights and the Law* 1.
- Philipsen S, Stamhuis EF, and De Jong WM, 'Legal Enclaves as a Test Environment for Innovative Products: Toward Legally Resilient Experimentation Policies' (2021) 15(4) *Regulation & Governance* 1128.
- Phillips PJ and O'Toole AJ, 'Comparison of Human and Computer Performance Across Face Recognition Experiments' (2014) 32(1) *Image and Vision Computing* 74.
- and others, 'Face Recognition Accuracy of Forensic Examiners, Superrecognizers, and Face Recognition Algorithms' (2018) 115 *Proceedings of the National Academy of Sciences* 6171.
- Pierce R, 'Machine Learning for Diagnosis and Treatment' (2018) 4 *European Data Protection Law Review* 340.
- Pietrogiovanni V, 'Deliveroo and Riders' Strikes: Discriminations in the Age of Algorithms' (2021) 7 *International Labour Rights Case Law* 317.
- Pinem AA and others, 'Trust and its Impact Towards Continuance of Use in Government-to-Business Online Service' (2018) 12(3–4) *Transforming Government: People, Process and Policy* 265–85.
- Pitruzzella G and Pollicino O, *Disinformation and Hate Speech* (Bocconi UP 2020).
- Pizzi M and others, 'AI for Humanitarian Action: Human Rights and Ethics' (2020) 102(913) *International Review of the Red Cross* 145.
- Poincare H, *Science and Hypothesis* (Scott 1905).

- Poklaski A, 'Toward an International Constitution of Patient Rights' (2016) 23 Indiana Journal of Global Legal Studies 893.
- Pollicino O, *Judicial Protection of Fundamental Rights Online: A Road Towards Digital Constitutionalism?* (Hart 2021).
- and De Gregorio G, 'Constitutional Law in the Algorithmic Society' in A Reichman and others (eds), *Constitutional Challenges in the Algorithmic Society* (CUP 2021) 3.
- Poncibo C and Zoboli L, 'Sandboxes and Consumer Protection: The European Perspective' (2020) 8 International Journal on Consumer Law and Practice 1.
- Ponta SE, Plate H, and Sabetta A, 'Detection, Assessment and Mitigation of Vulnerabilities in Open Source Dependencies' (2020) 25 Empirical Software Engineering 3175.
- Poole E, 'Fighting Back Against Non-Consensual Pornography' (2015) 49 University of San Francisco Law Review 181.
- Poort J and Zuiderveen Borgesius F, 'Does Everyone Have a Price? Understanding People's Attitude Towards Online and Offline Price Discrimination' (2019) 8(1) Internet Policy Review 1.
- Popova M, 'Reading Out of Context: Pornographic Deepfakes, Celebrity and Intimacy' (2019) 7 Porn Studies 367.
- Prakash P and others, 'Privacy Preserving Facial Recognition against Model Inversion Attacks' in IEEE Global Communications Conference (2020).
- Prasad S and Aravindakshan S, 'Playing Catch Up—Privacy Regimes in South Asia' (2021) 25(1) International Journal of Human Rights 79.
- Prifti K, Stamhuis E, and Heine K, 'Digging into the Accountability Gap: Operator's Civil Liability in Healthcare AI-systems' in B Custers and E Fosch-Villaronga (eds), *Law and Artificial Intelligence: Information Technology and Law Series* (TMC Asser Press 2022) 279.
- Prince A and Schwarcz D, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data' (2020) 105 Iowa Law Review 1257.
- Purshouse J and Campbell L, 'Automated Facial Recognition and Policing: A Bridge Too Far?' (2021) 42(2) Legal Studies 209.
- Purushothaman A and Palaniswamy S, 'Development of Smart Home Using Gesture Recognition for Elderly and Disabled' (2020) 17 Journal of Computational and Theoretical Nanoscience 177.
- Quinn P, 'Crisis Communication in Public Health Emergencies: The Limits of "Legal Control" and the Risks for Harmful Outcomes in a Digital Age' (2018) 14 Life Sciences, Society and Policy 4.
- Quintavalla A and Heine K, 'Priorities and Human Rights' (2019) 23(4) International Journal of Human Rights 679.
- Rachlinski JJ and others, 'Does Unconscious Racial Bias Affect Trial Judges?' (2009) Cornell Law Faculty Publications Paper 3/2009 786.
- Rachovitsa A and Johann N, 'The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case' (2022) 22 Human Rights Law Review 1.
- Rademacher T, 'Artificial Intelligence and Law Enforcement' in T Wischmeyer and T Rademacher (eds), *Regulating Artificial Intelligence* (Springer 2020) 228.
- Ragas J, 'La batalla por los rostros: el sistema de reconocimiento facial en el contexto del "estallido social" chileno' (2020) 14 Meridional Revista Chilena de Estudios Latinoamericanos 247.
- Rähme B, 'Artificial Intelligence and Religion: Between Existing AI and Grand Narratives' (2021) 17(4) Journal of Objects, Art and Belief 1.
- Rainey B, Wicks E, and Ovey C, *The European Convention on Human Rights* (OUP 2017).

- Ramanathan N, Chellappa R, and Biswas S, 'Computational Methods for Modeling Facial Aging: A Survey' (2009) 20(3) *Journal of Visual Languages and Computing* 131.
- Ranchordas S, 'Citizens as Consumers in the Data Economy: The Case of Smart Cities' (2018) 7 *Journal of European Consumer and Market Law* 154.
- , 'Nudging Citizens through Technology in Smart Cities' (2020) 34(3) *International Review of Law, Computers & Technology* 254.
- , and Scarella L, 'Automated Government for Vulnerable Citizens' (2022) 72 *William & Mary Bill of Rights Journal* 373.
- , and Schuurmans Y, 'Outsourcing the Welfare State: The Role of Private Actors in Welfare Fraud Investigations' (2020) 7 *European Journal of Comparative Law and Governance* 5.
- , and Van 't Schip M, 'Future-Proofing Legislation for the Digital Age' in S Ranchordas and Y Roznai (eds), *Time, Law, and Change: An Interdisciplinary Study* (Hart 2020) 347.
- Raso F and others, *Artificial Intelligence & Human Rights: Opportunities and Risks* (The Berkman Klein Center for Internet & Society at Harvard University 2018).
- Raz J, 'Promises in Morality and Law' (1982) 95 *Harvard Law Review* 916.
- Reddy S, Fox J, and Purohit M, 'Artificial Intelligence-Enabled Healthcare Delivery' (2019) 112(1) *Journal of the Royal Society of Medicine* 22.
- Reed R, 'AI in Religion, AI for Religion, AI and Religion: Towards a Theory of Religious Studies and Artificial Intelligence' (2021) 12(6) *Religions* 401.
- Regehr C and Alaggia R, 'Perspectives of Justice for Victims of Sexual Violence' (2006) 1(1) *Victims & Offenders* 33.
- Regehr K, Birze A, and Regehr C, 'Technology Facilitated Re-Victimization: How Video Evidence of Sexual Violence Contributes to Mediated Cycles of Abuse' (2021) 18(4) *Crime Media Culture* 1.
- Reid S, 'The Deepfake Dilemma: Reconciling Privacy and First Amendment Protections' (2021) 23(1) *University of Pennsylvania Journal of Constitutional Law* 209.
- Rengel A, *Privacy in the 21st Century* (Brill 2013).
- Reutter L, 'Constraining Context: Situating Datafication in Public Administration' (2022) 24(4) *New Media & Society* 903.
- Richardson BJ, *Socially Responsible Investment Law: Regulating the Unseen Polluters* (OUP 2008).
- Richardson R, Schultz JM, and Crawford K, 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice' (2019) 94 *New York University Law Review Online* 15.
- Ridgway JP and others, 'Machine Learning and Clinical Informatics for Improving HIV Care Continuum Outcomes' (2021) 3 *Current HIV/AIDS Reports* 229.
- Rigby MJ, 'Ethical Dimensions of Using Artificial Intelligence in Health Care' (2019) 21(2) *AMA Journal of Ethics* 121.
- Risse M, 'Human Rights and Artificial Intelligence: An Urgently Needed Agenda' (2019) 41 *Human Rights Quarterly* 7.
- , 'The Fourth Generation of Human Rights: Epistemic Rights in Digital Lifeworlds' 2021 (8) *Moral Philosophy and Politics* 351.
- , and Livingston S, 'The Future Impact of Artificial Intelligence on Humans and Human Rights' (2019) 33(2) *Ethics & International Affairs* 141.
- Rivers J, 'Proportionality and Discretion in International and European Law' in N Tsagourias (ed), *Transnational Constitutionalism: International and European Perspectives* (CUP 2010) 107.
- Roberts ST, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale UP 2019).

- Robertson AH, *Collected Edition of the 'Travaux Préparatoires' of the European Convention on Human Rights = Recueil des Travaux Préparatoires de la Convention Européenne des Droits de l'Homme* (Nijhoff 1975).
- Rochet J-C and Tirole J, 'Two-sided Markets: A Progress Report' (2006) 37(3) *The RAND Journal of Economics* 645.
- Rodrigues R, 'Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities' (2020) 4 *Journal of Responsible Technology* 100005.
- Rodriguez NS and Hernandez T, 'Dibs on that Sexy Piece of Ass: Hegemonic Masculinity on TFM Girls Instagram Account' (2018) 4(1) *Social Media & Society* 1.
- Roe Smith M and Marx L (eds), *Does Technology Drive History?: The Dilemma of Technological Determinism* (MIT Press 1994).
- Roessler B, 'Three Dimensions of Privacy' in B van der Sloot and A de Groot (eds), *The Handbook of Privacy Studies: An Interdisciplinary Introduction* (AUP 2018) 137.
- Rolnik R, 'The Human Right to Adequate Housing' in FZ Giustiniani and others (eds), *Routledge Handbook of Human Rights and Disasters* (Routledge 2018) 180.
- Ronsinet X and others, *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment: Appendix I—In-Depth Study on the Use of AI in Judicial Systems, Notably AI Applications Processing Judicial Decisions and Data* (CEPEJ 2018).
- Rosenblat A and Stark L, 'Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers' (2016) 10 *International Journal of Communication* 3758.
- and others, 'Discriminating Tastes: Customer Ratings as Vehicles for Bias' (2016) *Data & Society* 1.
- Rosol M and Blue G, 'From the Smart City to Urban Justice in a Digital Age' (2022) 26(4) *City* 684.
- Rudin C, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1(5) *Nature Machine Intelligence* 1.
- Rudkin T, 'Things Get Serious: Defining Defamation' (2014) 25(6) *Entertainment Law Review* 201.
- Russel S and Norvig P, *Artificial Intelligence: A Modern Approach* (4th edn, Pearson 2021).
- Ryberg J, *Domstolens blinde øje: Om betydningen af ubevidste biases i retssystemet* (Djof 2016).
- , 'Risk-Based Sentencing and Predictive Accuracy' (2020) 23 *Ethical Theory and Moral Practice* 271.
- Sabu KM and Kumar TKM, 'Predictive Analytics in Agriculture: Forecasting Prices of Areca Nuts in Kerala' (2020) 171 *Procedia Computer Science* 699.
- Šadl U and Olsen HP, 'Can Quantitative Methods Complement Doctrinal Legal Studies? Using Citation Network and Corpus Linguistic Analysis to Understand International Courts' (2017) 30 *Leiden Journal of International Law* 327.
- Sahni S, 'SURVEy: Techniques for Aging Problems in Face Recognition' (2014) 4(2) *MIT International Journal of Computer Science and Information Technology* 1.
- Salathé M, Wiegand T, and Wenzel M, 'Focus Group on Artificial Intelligence for Health' (2018) World Health Organization 3.
- Salvini P and others, 'An Investigation on Legal Regulations for Robot Deployment in Urban Areas: A Focus on Italian Law' (2021) 24 *Advanced Robotics* 1901–17.
- Sampani C, 'Online Dispute Resolution in E-Commerce: Is Consensus in Regulation UNCITRAL's Utopian Idea or a Realistic Ambition?' (2021) 30(3) *Information & Communications Technology Law* 235.
- Samuelson P, 'Privacy as Intellectual Property?' (2000) 52 *Stanford Law Review* 1125.
- Sartor G, 'Artificial Intelligence and Human Rights: Between Law and Ethics' (2020) 27 *Maastricht Journal of European and Comparative Law* 705.

- and Lagioia F, ‘The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence’ (EU Parliament PE641.530, 2020).
- Sayers D, ‘Article 47(2): Everyone is Entitled to a Fair and Public Hearing Within a Reasonable Time by an Independent and Impartial Tribunal Previously Established by Law. Everyone Shall Have the Possibility of Being Advised, Defended and Represented’ in S Peers and others (eds), *The EU Charter of Fundamental Rights: a Commentary* (Hart 2015).
- Scassa T, ‘A Human Rights-Based Approach to Data Protection in Canada’ in E Dubois and F Martin-Bariteau (eds), *Citizenship in a Connected Canada: A Research and Policy Agenda* (University of Ottawa Press 2020) 183.
- Schauer F, ‘Fear, Risk and the First Amendment: Unravelling the “Chilling Effect”’ (1978) Boston University Law Review 689.
- , ‘Giving Reasons’ (1995) 47 Stanford Law Review 633.
- , *Thinking Like a Lawyer* (Harvard UP 2009).
- Schaupp S, ‘Algorithmic Integration and Precarious (Dis)Obedience: On the Co-Constitution of Migration Regime and Workplace Regime in Digitalised Manufacturing and Logistics’ (2021) 36 Work, Employment and Society 310.
- Schermer B, ‘Ambient Intelligence, Persoonsgegevens En Consumentenbescherming’ (ECP Platform voor de InformatieSamenleving 2008).
- Scheuerman MK, Paul JM, and Brubaker JR, ‘How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services’ (2019) 3 Proceedings of the ACM on Human-Computer Interaction 1.
- Schlosser S, Toninelli D, and Cameletti M, ‘Comparing Methods to Collect and Geolocate Tweets in Great Britain’ (2021) 7 Journal of Open Innovation: Technology, Market, and Complexity 44.
- Schmidt P, Biessmann F, and Teubner T, ‘Transparency and Trust in Artificial Intelligence Systems’ (2020) 29(4) Journal of Decision Systems 260–78.
- Schmitz AJ, ‘Secret Consumer Scores and Segmentations: Separating Haves from Have-Nots’ (2014) Michigan State Law Review 1411.
- Schnieble CO, Elger BS, and Shaw DM, ‘Google’s Project Nightingale Highlights the Necessity of Data Science Ethics Review’ (2020) 12(3) EMBO Molecular Medicine 1.
- Schoenholtz AI, Ramji-Nogales J, and Schrang PG, ‘Refugee Roulette: Disparities in Asylum Adjudication’ (2007) 60 Stanford Law Review 295.
- Schragger R, *City Power: Urban Governance in a Global Age* (OUP 2016).
- Schüller B, ‘The Definition of Consumers in EU Consumer Law’ (2012) European Consumer Protection 123.
- Schütte B, Majewski L, and Havu K, ‘Damages Liability for Harm Caused by Artificial Intelligence—EU Law in Flux’ (2021) Helsinki Legal Studies Research Paper No 69.
- Schwalbe N and Wahl B, ‘Artificial Intelligence and the Future of Global Health’ (2020) 395 *The Lancet* 1579.
- Schwartz R and others, ‘Green AI’ (2020) 63(12) Communications of the ACM 54.
- Schwebel E, ‘The International Convention on the Elimination of all Forms of Racial Discrimination’ (1966) 15 International and Comparative Law Quarterly 996.
- Scott-Hayward CS, Fradella HF, and Fischer RG, ‘Does Privacy Require Secrecy: Societal Expectations of Privacy in the Digital Age’ (2015) 43 American Journal of Criminal Law 19.
- Searle JR, ‘Minds, Brains and Programs’ (1980) 3 Behavioral and Brain Sciences 417.
- Sears AM, ‘Algorithmic Speech and Freedom of Expression’ (2020) 53 Vanderbilt Journal of Transnational Law 1327.

- Selbst AD and Barocas S, 'The Intuitive Appeal of Explainable Machines' (2018) 87 Fordham Law Review 1085.
- Selinger E and Hartzog W, 'The Inconsentability of Facial Surveillance' (2020) 66 Loy Law Review 101.
- Seng S and others, 'A First Look into Users' Perceptions of Facial Recognition in the Physical World' (2021) 105 Computers & Security 1.
- Šepčec M and Lango M, 'Virtual Revenge Pornography As A New Online Threat To Sexual Integrity' (2020) 15 Balkan Social Science Review 117.
- Sepúlveda Carmona M, *The Nature of the Obligations Under the International Covenant on Economic, Social and Cultural Rights* (Intersentia 2003).
- Shaban-Nejad A, Michalowski M, and Buckeridge DL, 'Health Intelligence: How Artificial Intelligence Transforms Population and Personalized Health' (2018) 1(2) NPJ Digital Medicine 53.
- Shah S, 'Detention and Trial' in D Moeck and others (eds), *International Human Rights Law* (OUP 2014) 257.
- Sharif M and others, 'Accessorize to a Crime' in ACM SIGSAC Conference on Computer and Communications Security (2016) 1533.
- Sharkey A, 'Autonomous Weapons Systems, Killer Robots and Human Dignity' (2019) 21 Ethics and Information Technology 75.
- Shaw J, 'Platform Real Estate: Theory and Practice of New Urban Real Estate Markets' (2020) 41 Urban Geography 1037.
- Shearer C, 'The CRISP-DM Model: The New Blueprint for Data Mining' (2000) 5 Journal of Data Warehousing 13.
- Sheir S and others, 'Algorithmic Impact Assessments. Building a Systematic Framework of Accountability for Algorithmic Decision Making' (Institute for the Future of Work 2021).
- Shelton D, 'Sources of Article 47 Rights' in S Peers and others (eds), *The EU Charter of Fundamental Rights: a Commentary* (Hart 2015).
- Shenkman C, Thakur D, and Llansó E, *Do You See What I See: Capabilities and Limits of Automated Multimedia Content Analysis* (CDT 2021).
- Shi P and others, *World Atlas of Natural Disaster Risk* (Springer 2015).
- Shoruzzaman M, Hossain MS, and Alhamid MF, 'Towards the Sustainable Development of Smart Cities through Mass Video Surveillance: A Response to the COVID-19 Pandemic' (2021) 64 Sustainable Cities and Society 102582.
- Shue H, *Basic Rights: Subsistence, Affluence, and American Foreign Policy* (Princeton UP 1980).
- Siatitsa I, 'Freedom of Assembly Under Attack: General and Indiscriminate Surveillance and Interference with Internet Communications' (2020) 102(913) International Review of the Red Cross 191.
- Siebers T, *Disability Theory* (University of Michigan Press 2008).
- Silver C, *Urban Flood Risk Management: Looking at Jakarta* (Routledge 2022).
- Singler B, 'An Introduction to Artificial Intelligence and Religion for the Religious Studies Scholar' (2017) 20(3) Journal of Implicit Religion 215.
- , "Blessed by the Algorithm": Theistic Conceptions of Artificial Intelligence in Online Discourse' (2020) 3 AI & Society 945.
- Sinharoy SS, Pittluck R, and Clasen T, 'Review of Drivers and Barriers of Water and Sanitation Policies for Urban Informal Settlements in Low-Income and Middle-Income Countries' (2019) 60 Utilities Policy 100957.
- Skinner-Thompson S, 'Agonistic Privacy & Equitable Democracy' (2021) 131 Yale Law Journal Forum 459.

- Skitka L, Mosier K, and Burdick M, 'Does Automation Bias Decision-Making?' (1999) 51(5) International Journal of Human-Computer Studies 991.
- Skopik F and Smith P (eds), *Smart Grid Security: Innovative Solutions for a Modernized Grid* (Syngress 2015).
- Slove DJ, 'Conceptualizing Privacy' (2002) 90 California Law Review 1087.
- Smith JK, *Robotic Persons* (Westbow 2021).
- Sokhi-Bulley B, 'The Optional Protocol to CEDAW: First Steps' (2006) 6 Human Rights Law Review 143.
- Solove D, 'Conceptualizing Privacy' (2002) 90 California Law Review 1087.
- , *Nothing to Hide: The False Tradeoff Between Privacy and Security* (1st edn, Yale UP 2011).
- Spagnolo A, 'What do Human Rights Really Say About the Use of Autonomous Weapons Systems for Law Enforcement Purposes?' in E Carpanelli and N Lazzerini (eds), *Use and Misuse of New Technologies: Contemporary Challenges in International and European Law* (Springer 2019) 55.
- Sparrow R, 'Killer Robots' (2007) 24(1) Journal of Applied Philosophy 62.
- Spencer D and others, 'I Think It's Re-Victimizing Victims Almost Every Time': Police Perceptions of Criminal Justice Responses to Sexual Violence' (2018) 26(2) Critical Criminology 189.
- Sperti A, 'The Impact of Information and Communication Revolution on Constitutional Courts' in Martin Belov (ed), *The IT Revolution and Its Impact on State, Constitutionalism and Public Law* (Bloomsbury 2021) 184.
- Spielmann D, 'Whither the Margin of Appreciation?' (2014) 67(1) Current Legal Problems 54.
- Spina Alí G, 'Intellectual Property and Human Rights: A Taxonomy of Their Interactions' (2020) 51 IIC: International Review of Intellectual Property and Competition Law 411.
- and Yu R, 'Artificial Intelligence between Transparency and Secrecy: From the EC Whitepaper to the AIA and Beyond' (2021) 12 European Journal of Law and Technology 1.
- Spivak R, 'Deepfakes': The Newest Way to Commit One of the Oldest Crimes' (2019) 3(2) Georgetown Law Technology Review 339.
- Sprague R, 'Welcome to the Machine: Privacy and Workplace Implications of Predictive Analytics' (2014) 21 Richmond Journal Law & Technology 1.
- Staben J, *Der Abschreckungseffekt auf die Grundrechtsausübung—Strukturen eines Verfassungsrechtlichen* (Mohr Siebeck 2016).
- Stahl T, 'Privacy in Public: A Democratic Defence' (2020) 7(1) Moral Philosophy and Politics 74.
- Stark L and Hutson J, 'Physiognomic Artificial Intelligence' (2022) 32 Fordham Intellectual Property, Media & Entertainment Law Journal 922.
- Steed R and Caliskan A, 'Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases' (2021) Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 701.
- Steele L and others, 'Human Rights and the Confinement of People Living with Dementia in Care Homes' (2020) 22(1) Health and Human Rights Journal 7.
- Steeves V, 'Reclaiming the Social Value of Privacy' in Ian Kerr, Valerie Steeves, and Carole Lucock (eds), *Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society* (OUP 2009) 191.
- Steffen W and others, 'Planetary Boundaries: Guiding Human Development on a Changing Planet' (2015) 347(6223) Science 1259855.

- Steponenaite VK and Valcke P, 'Judicial Analytics on Trial: An Assessment of Legal Analytics in Judicial Systems in Light of the Right to a Fair Trial' (2020) 27(6) Maastricht Journal of European and Comparative Law 759.
- Stern A, 'Innovation under Regulatory Uncertainty: Evidence from Medical Technology' (2018) 145 Journal of Public Economics 181.
- Stevenson MT, 'Assessing Risk Assessment in Action' (2018) 103 Minnesota Law Review 303.
- Stewart P and Stuhmcke A, 'Judicial Analytics and Australian Courts: A Call for National Ethical Guidelines?' (2019) 45(2) Alternative Law Journal 1.
- Stewart RB, 'Administrative Law in the Twenty-first Century' (2003) 78 New York University Law Review 437.
- Stiglitz JE, 'Transparency in Government' in World Bank (ed), *The Right to Tell: The Role of Mass Media in Economic Development* (World Bank 2002) 27.
- Strikwerda L, 'Predictive Policing: The Risks Associated with Risk Assessment' (2021) 94 The Police Journal 422.
- Strubell E, Ganesh A, and McCallum A, 'Energy and Policy Considerations for Deep Learning in NLP' (2019) arXiv:1906.02243.
- Sturm S, 'Second Generation Employment Discrimination: A Structural Approach' (2001) 101 Columbia Law Review 458.
- Suleimenova D, Bell D, and Groen D, 'A Generalized Simulation Development Approach for Predicting Refugee Destinations' (2017) 7 Scientific Reports 13377.
- Sun Z and others, 'A Review of Earth Artificial Intelligence' (2022) 159 Computers & Geosciences 105034.
- Sunstein CR, #*Republic: Divided Democracy in the Age of Social Media* (Princeton UP 2017).
- , 'Sludge and Ordeals' (2018) 68 Duke Law Journal 1843.
- , 'Governing by Algorithm? No Noise and (Potentially) Less Bias' (2022) 71 Duke Law Journal 1175.
- Susser D, 'Predictive Policing and the Ethics of Preemption' in B Jones and E Mendieta (eds), *The Ethics of Policing: New Perspectives on Law Enforcement* (NYU Press 2021) 268.
- , Roessler B, and Nissenbaum H, 'Technology, Autonomy, and Manipulation' (2019) 8(2) Internet Policy Review 1.
- Susskind R, *Online Courts and the Future of Justice* (OUP 2019).
- Suzor NP, *Lawless: The Secret Rules That Govern Our Digital Lives* (CUP 2019).
- Swanson G, 'Non-Autonomous Artificial Intelligence Programs and Products Liability: How New AI Products Challenge Existing Liability Models and Pose New Financial Burdens' (2019) 42 Seattle University Law Review 1203.
- Swanson M and others, 'Contingency Planning Guide for Federal Information Systems' (National Institute of Standards and Technology 2010).
- Takagi H and others, 'Projection of Coastal Floods in 2050 Jakarta' (2016) 17 Urban Climate 135.
- Tamburini G, *Etica delle Macchine* (Carocci 2020).
- Tavani HT, 'Philosophical Theories of Privacy: Implications for an Adequate Online Privacy Policy' (2007) 38 Metaphilosophy 1.
- Taylor L, Van der Sloot B, and Floridi L, *Group Privacy: New Challenges of Data Technologies* (Springer 2016).
- Terhörst P and others, 'On Soft-Biometric Information Stored in Biometric Face Embeddings' (2021) 4 IEEE Transactions on Biometrics, Behavior, and Identity Science 519.
- Terzidou K, 'The Use of Artificial Intelligence in the Judiciary and Its Compliance with the Right to a Fair Trial' (2022) 31(3) Journal of Judicial Administration 154.

- Teubner G, 'Digital Personhood? The Status of Autonomous Software Agents in Private Law' (2018) 106 *Ancilla Iuris* 35.
- Theodorou A, Wortham RH, and Bryson JJ, 'Designing and Implementing Transparency for Real Time Inspection of Autonomous Robots' (2017) 29(3) *Connection Science* 230.
- Thompson KA, 'Countenancing Employment Discrimination: Facial Recognition in Background Checks' (2020) 8 *Texas A&M Law Review* 63.
- Thornberry P, *The International Convention on the Elimination of All Forms of Racial Discrimination* (OUP 2016).
- Thorun C and Diels J, 'Consumer Protection Technologies: An Investigation into the Potentials of New Digital Technologies for Consumer Policy' (2020) 43 *Journal of Consumer Policy* 182.
- Todoli-Signes A, 'Making Algorithms Safe for Workers: Occupational Risks associated with Work Managed by Artificial Intelligence' (2021) 27 *Transfer* 433.
- Tomasev N and others, 'Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities', Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (2021) 254.
- Tomlinson J, Maxwell J, and Welsh A, 'Discrimination in Digital Immigration Status' (2022) 42(2) *Legal Studies* 315.
- Tosun J, 'How the EU Handles Uncertain Risks: Understanding the Role of the Precautionary Principle' (2013) 20 *Journal of European Public Policy* 1517.
- Trewin S, 'AI Fairness for People with Disabilities: Point of View' (2018) arXiv:1811.10670.
- Tünsmeyer V, *Repatriation of Sacred Indigenous Cultural Heritage and the Law: Lessons from the United States and Canada* (Springer 2021).
- Turing AM, 'Computing Machinery and Intelligence' (1950) 49 *Mind* 433.
- Turley J, 'Anonymity, Obscurity, and Technology: Reconsidering Privacy in the Age of Biometrics' (2020) 100(6) *Boston University Law Review* 2179.
- Turman-Bryant N and others, 'Toilet Alarms: A Novel Application of Latrine Sensors and Machine Learning for Optimizing Sanitation Services in Informal Settlements' (2020) 5 *Development Engineering* 100052.
- Turner J, *Robot Rules: Regulating Artificial Intelligence* (Palgrave 2019).
- Tutt A, 'An FDA for Algorithms' (2017) 69(1) *Administrative Law Review* 83.
- Uchida CD and Swatt ML, 'Operation LASER and the Effectiveness of Hotspot Patrol: A Panel Analysis' (2013) 16 *Police Quarterly* 287.
- Umoja Noble S, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press 2018).
- Vaele M and Brass I, 'Administration by Algorithm? Public Management Meets Public Sector Machine Learning' in K Yeung and M Lodge (eds), *Algorithmic Regulation* (OUP 2019) 123–27.
- Valentine S, 'Impoverished Algorithms: Misguided Governments, Flawed Technologies, and Social Control' (2019) 46(2) *Fordham Urban Legal Journal* 364.
- Van Aalst MK and others, 'The Impacts of Climate Change on the Risk of Natural Disasters' (2006) 30 *Disasters* 5.
- Van Bekkum V and Zuiderveen Borgesius F, 'Digital Welfare Fraud Detection and the Dutch SyRI Judgment' (2021) 23(4) *European Journal of Social Security* 323.
- Van Brakel R, 'How to Watch the Watchers? Democratic Oversight of Algorithmic Police Surveillance in Belgium' (2021) 19 *Surveillance & Society* 228.
- Van de Hoven J and others, 'Towards a Digital Ecosystem of Trust: Ethical, Legal and Societal Implications' (2021) *Opinio Juris in Comparatione* 131–56.

- Van der Nagel E, 'Verifying Images: Deepfakes, Control, and Consent' (2020) 7(4) *Porn Studies* 424.
- Van der Sloot B, *Privacy as Virtue* (Intersentia 2017).
- , 'A New Approach to the Right to Privacy, or How the European Court of Human Rights Embraced the Non-Domination Principle' (2018) 34(3) *Computer Law & Security Review* 539.
- , 'The Right to Be Let Alone by Oneself: Narrative and Identity in a Data-Driven Environment' (2021) 13(1) *Law, Innovation and Technology* 223.
- Van Dijk P and others, *Theory and Practice of the European Convention on Human Rights* (Intersentia 2006).
- Van Hartkamp M and others, 'Artificial Intelligence in Clinical Health Care Applications: Viewpoint' (2019) 8(2) *Interactive Journal of Medical Research* 1.
- Van Natta M and others, 'The Rise and Regulation of Thermal Facial Recognition Technology during the COVID-19 Pandemic' (2020) 7(1) *Journal of Law and the Biosciences* 1.
- Van Noorden R, 'The Ethical Questions that Haunt Facial-Recognition Research' (2020) 587(7834) *Nature* 354.
- Van Veen C and Cath C, 'Artificial Intelligence: What's Human Rights Got To Do With It?' (2018) 14 *Data & Society*.
- Vanolo A, 'Cities and the Politics of Gamification' (2018) 74 *Cities* 320.
- Vasudeva A, Sheikh NA, and Sahu S, 'International Classification of Functioning, Disability, and Health Augmented by Telemedicine and Artificial Intelligence for Assessment of Functional Disability' (2021) 10(10) *Journal of Family Medicine and Primary Care* 3535.
- Veeder, VV, 'The History and Theory of the Law of Defamation: II' (1904) 4(1) *Columbia Law Review* 33.
- Véliz C, *Privacy Is Power* (Melville House 2021).
- Velman M, *Understanding Consciousness* (Routledge 2000).
- Verma M, 'Lexical Analysis of Religious Texts using Text Mining and Machine Learning Tools' (2017) 168(8) *International Journal of Computer Applications* 39.
- Vermeule A, 'Security and Liberty: Critiques of the Trade-off Thesis' in David Jenkins and others (eds), *The Long Decade: How 9/11 Changed the Law* (OUP 2014) 31.
- Vetzo M and Gerards J, 'Algoritme-Gedreven Technologieën en Grondrechten' (2019) 21 *Computerrecht* 10.
- Vihalemm T and Keller M, 'Consumers, Citizens or Citizen-consumers? Domestic Users in the Process of Estonian Electricity Market Liberalization' (2016) 13 *Energy Research & Social Science* 38.
- Viljanen M and Parviainen H, 'AI Applications and Regulation: Mapping the Regulatory Strata' (2022) 3 *Frontiers* 1.
- Vinuesa R and others, 'The Role of Artificial Intelligence in Achieving the Sustainable Development Goals' (2020) 11(1) *Nature Communications* 1.
- Von Arnauld A, Von der Decken K, and Susi M (eds), *The Cambridge Handbook of New Human Rights: Recognition, Novelty, Rhetoric* (1st edn, CUP 2020).
- Voorwinden A, 'The Privatised City: Technology and Public-Private Partnerships in the Smart City' (2021) 13(3) *Law, Innovation, and Technology* 439.
- Wachter S, 'Data Protection in the Age of Big Data' (2019) 2(1) *Nature Electronics* 6.
- , Mittelstadt B, and Floridi L, 'Transparent, Explainable, and Accountable AI for Robotics' (2017) 2 *Science Robotics* 1.
- , Mittelstadt B, and Floridi L, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7(2) *International Data Privacy Law* 76.

- , Mittelstadt B, and Russell C, 'Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law' (2021) 123 West Virginia Law Review 735.
- Wagner B, Kettemann MC, and Vieth K, *Research Handbook on Human Rights and Digital Technology: Global Politics, Law and International Relations* (Edward Elgar 2019).
- Wagner TL and Blewer A, ““The Word Real Is No Longer Real”: Deepfakes, Gender, and the Challenges of AI-Altered Video’ (2019) 3 Open Information Science 36.
- Wahl B and others, ‘Artificial Intelligence (AI) and Global Health: How Can AI Contribute to Health in Resource-Poor Settings?’ (2018) 3(4) BMJ Global Health 1.
- Wahl F and others, ‘Mobile Sensing and Support for People with Depression: A Pilot Trial in the Wild’ (2016) 4(3) JMIR mHealth and uHealth 5960.
- Waldman AE, ‘Power, Process, and Automated Decision-Making’ (2019) 88 Fordham Law Review 613.
- , *Industry Unbound* (CUP 2021).
- Walters R and Novak M, *Cyber Security, Artificial Intelligence, Data Protection & the Law* (Springer 2021).
- Wang B, Loo B, and Huang G, ‘Becoming Smarter through Smart City Pilot Projects: Experiences and Lessons from China since 2013’ (2021) 29(4) Journal of Urban Technology 3.
- Wang J and others, ‘Unobtrusive Health Monitoring in Private Spaces: The Smart Home’ (2021) 21 Sensors 864.
- Wang Y and Kosinski M, ‘Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images’ (2018) 114(2) Journal of Personality and Social Psychology 246.
- Ward BW and Schiller JS, ‘Prevalence of Multiple Chronic Conditions Among US Adults: Estimates From the National Health Interview Survey, 2010’ (2013) 10 Preventing Chronic Disease.
- Warren SD and Brandeis LD, ‘The Right to Privacy’ (1890) 4(5) Harvard Law Review 193, 195.
- Washington AL, ‘How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate’ (2018) 17 Colorado Technology Law Journal 131.
- Weber RH, ‘Accountability in the Internet of Things’ (2011) 27(2) Computer Law & Security Review 133.
- Weil D, *The Fissured Workplace* (Harvard UP 2014).
- Weinberg L, ‘Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches’ (2022) 74 Journal of Artificial Intelligence Research 75.
- Weiss AG, ‘Privacy and Intimacy: Apart and a Part’ (1987) 27 Journal of Humanist Psychology 1.
- Weis P, *The Refugee Convention, 1951: The Travaux Préparatoires Analysed with a Commentary by Dr Paul Weis* (1st edn, CUP 1995).
- Weissbrodt D and Kruger M, ‘Norms on the Responsibilities of Transnational Corporations and Other Business Enterprises with Regard to Human Rights’ (2003) 97 American Journal of International Law 901.
- Weng Y-H and others, ‘Intersection of “Tokku” Special Zone, Robots, and the Law: A Case Study on Legal Impacts to Humanoid Robots’ (2015) 7 International Journal of Social Robotics 841.
- Westerlund M, ‘The Emergence of Deepfake Technology: A Review’ (2019) 9 Technology Innovation Management Review 40.
- Westin A, *Privacy and Freedom* (Ig 1967).

- , ‘Social and Political Dimensions of Privacy’ (2003) 59 *Journal of Social Issues* 431.
- Wettstein F, ‘Normativity, Ethics, and the UN Guiding Principles on Business and Human Rights: A Critical Assessment’ (2015) 14 *Journal of Human Rights* 162.
- , ‘From Side Show to Main Act: Can Business and Human Rights Save Corporate Responsibility’ in D Baumann-Pauly and J Nolan (eds), *Business and Human Rights: From Principles to Practice* (Routledge 2016) 78.
- , *Business and Human Rights: Ethical, Legal, and Managerial Perspectives* (CUP 2022).
- Wetzling T and Vieth K, ‘Legal Safeguards and Oversight Innovations for Bulk Surveillance: An International Comparative Analysis’ in LA Viola and P Laidler (eds), *Trust and Transparency in an Age of Surveillance* (Routledge 2021) 145.
- Wexler R, ‘Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System’ (2018) 70 *Stanford Law Review* 1349.
- Whittaker M and others, ‘Disability, Bias, and AI’ (AI Now Institute 2019).
- Widlak A, Van Eck M, and Peeters R, ‘Towards Principles of Good Digital Administration’ in M Schuilenberg and R Peeters (eds), *The Algorithmic Society: Technology, Power, and Knowledge* (Routledge 2020) 67.
- Wiemann M, Meidert N, and Weibel A, ‘“Good” and “Bad” Control in Public Administration: The Impact of Performance Evaluation Systems on Employees’ Trust in the Employer’ (2019) 48 *Public Personnel Management* 283.
- Wilkinson PHC, ‘The Legal Implications of Sexual Orientation-Detecting Facial Recognition Technology’ (2021) 20 *Dukeminier Awards Journal* 301.
- Williams JC and others, ‘Colourblind Algorithms: Racism in the Era of Covid 19’ (2020) 112(5) *Journal of the National Medical Association* 550.
- Wilson R and Land M, ‘Hate Speech on Social Media: Content Moderation in Context’ (2021) 52 *Connecticut Law Review* 1029.
- Winner L, *The Whale and the Reactor: A Search for Limits in an Age of High Technology* (University of Chicago Press 1986).
- Winter E, ‘The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law’ (2022) 27 *Journal of Conflict and Security Law* 1.
- Wiseman L and others, ‘Farmers and their Data: An Examination of Farmers’ Reluctance to Share their Data through the Lens of the Law Impacting Smart Farming’ (2019) 90–91 *NJA\|S Wageningen Journal of Life* 100301.
- Witzleb N and others, ‘An Overview of Emerging Challenges in Privacy Law’ in N Witzleb and others (eds), *Emerging Challenges in Privacy Law: Comparative Perspectives* (CUP 2014) 1.
- Wolfert S and others, ‘Big Data in Smart Farming’ (2017) 153 *Agric Systems* 69.
- Wolfson J, ‘The Expanding Scope of Human Rights in a Technological World—Using the InterAmerican Court of Human Rights to Establish a Minimum Data Protection Standard Across Latin America’ (2017) 48(3) *University of Miami Inter-American Law Review* 188.
- Wolswinkel J, *Willekeur of Algoritme? Laveren tussen Analoog en Digitaal Bestuursrecht* (Tilburg University 2020).
- Wood AJ, *Algorithmic Management: Consequences for Work Organisation and Working Conditions* (JRC Working Papers Series on Labour, Education and Technology 2021/07; European Commission 2021) 2021.
- and others, ‘Good Gig, Bad Gig: Autonomy and Algorithmic Control in the Global Gig Economy’ (2019) 33 *Work, Employment and Society* 56.
- Wright D and De Hert P, ‘Introduction to Privacy Impact Assessment’ in D Wright and P De Hert (eds), *Privacy Impact Assessment* (Springer 2012) 3–32.

- and Raab C, 'Privacy Principles, Risks and Harms' (2014) 28 International Review of Law, Computers and Technology 277.
- Wright N (ed), *AI, China, Russia, and the Global Order: Technological, Political, Global, and Creative Perspectives* (Report of the United States Department of Defense 2018).
- Wu T, 'Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems' (2019) 119 Columbia Law Review 2001.
- Wu X and Zhang X, 'Automated Inference on Criminality Using Face Images' (2016) arXiv 1611.04135.
- and Zhang X, 'Responses to Critiques on Machine Learning of Criminality Perceptions' (2016) arXiv 1611.04135.
- Wulf J and Seizov O, '"Please Understand We Cannot Provide Further Information": Evaluating Content and Transparency of GDPR-mandated AI Disclosures' (2022) *AI & Society*.
- Wuyts D, 'The Product Liability Directive: More than two Decades of Defective Products in Europe' (2014) 5 *Journal of European Tort Law* 21.
- Xenidis R, 'Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience' (2021) 27 *Maastricht Journal of European and Comparative Law* 736.
- Xu Y and Ma H, 'Research and Implementation of the Text Matching Algorithm in the Field of Housing Law and Policy Based on Deep Learning' (2021) *Complexity*.
- Yam J and Skorburg JA, 'From Human Resources to Human Rights: Impact Assessments for Hiring Algorithms' (2021) 23 *Ethics and Information Technology* 611.
- Yang X, 'Accelerated Move for AI Education in China' (2019) 2(3) *ECNU Review of Education* 347.
- Yeung K, 'Recommendation of the Council on Artificial Intelligence (OECD)' (2020) 59 *International Legal Materials* 27.
- , Howes A, and Pogrebna G, 'AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing' in Markus D Dubber and others (eds), *Oxford Handbook of AI Ethics* (OUP 2019) 87.
- and Bygrave LA, 'Demystifying the Modernized European Data Protection Regime: Cross-disciplinary Insights from Legal and Regulatory Governance Scholarship' (2021) 16 *Regulation & Governance* 10.
- Yu R and Spina Ali G, 'What's Inside the Black Box? AI Challenges for Lawyers and Researchers' (2019) 19 *Legal Information Management* 2.
- Zalnieriute M, 'Burning Bridges: The Automated Facial Recognition Technology and Public State Surveillance in the Modern State' (2021) 22 *Columbia Science & Technology Law Review* 284.
- Zarsky TZ, 'Understanding Discrimination in the Scored Society' (2014) 89 *Washington Law Review* 1375.
- , 'Incompatible: The GDPR in the Age of Big Data' (2016) 47 *Seton Hall Law Review* 995.
- , 'The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making' (2016) 41 *Science, Technology & Human Values* 1.
- Zhang L and others, 'Gender Biases in Estimation of Others' Pain' (2021) 22(9) *The Journal of Pain* 1048.
- Zhang Z, Chen Z, and Xu L, 'Artificial Intelligence and Moral Dilemmas: Perception of Ethical Decision-Making in AI' (2022) 101 *Journal of Experimental Social Psychology* 104327.

- Zhou J and others, 'Sensor-Array Optimization Based on Time-Series Data Analytics for Sanitation-Related Malodor Detection' (2020) 14 IEEE Transactions on Biomedical Circuits and Systems 705.
- Zuboff S, 'Big Other: Surveillance Capitalism and the Prospects of an Information Civilization' (2015) 30 Journal of Information Technology 75.
- , *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power* (Profile Books 2019).
- Zuiderveen Borgesius F, 'Discrimination, Artificial Intelligence and Algorithmic Decision-Making' (Directorate General of Democracy of the Council of Europe 2018) 2018.
- , 'Strengthening Legal Protection Against Discrimination by Algorithms and Artificial Intelligence' (2020) 24 International Journal of Human Rights 1572.
- and Poort J, 'Online Price Discrimination and EU Data Privacy Law' (2017) 40 Journal of Consumer Policy 347.
- Zuriek E, 'Theorizing Surveillance' in D Lyon (ed), *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination* (1st edn, Routledge 2002) 32.
- Zwart T, 'More Human Rights than Court: Why the Legitimacy of the European Court of Human Rights is in Need of Repair and How It Can Be Done' in S Flogaitis, J Fraser, and T Zwart (eds), *The European Court of Human Rights and Its Discontents: Turning Criticism into Strength* (Edward Elgar 2013) 71.

Index

For the benefit of digital users, indexed terms that span two pages (e.g., 52–53) may, on occasion, appear on only one of those pages.

Note: Entries pertaining only to specific countries/regions, particularly European Union, United Kingdom and United States have only been listed as sub-entries under the respective places.

- 4Chan 543–44
5G networks 430
- Aarhus Convention 295n.5
abuse of AI *see* misuse and abuse of AI
accountability 4, 13–14
 AI decision-making limits 494–95, 496–97
 asylum rights 316–17, 325, 336, 338
 common risks and benefits 450–51, 453–54,
 456–57
 consumer protection rights 421–22
 corporate 521
 data analytics in justice system and
 healthcare 286–87
 fair trial rights 104–18
 housing rights 360
 law enforcement and criminal justice 48
 liberty and security rights 56, 60
 modelling 29–30
 privacy and personal data protection 127–28,
 133–34
 religious freedom rights 65–66, 70, 71–72
 women’s rights and deepfake
 pornography 243
- accuracy 28, 29–30
 asylum rights 323
effective remedy right and ADM 306
freedom of assembly: biometrics and
 surveillance 95
gender-based discrimination 210
healthy environment right 429
human rights risk assessment 38
liberty and security rights 56–57
predictive 420–21
privacy and personal data protection 126,
 130–31, 134–35
privacy, personal data protection and
 FRT 137–38
see also inaccuracy
adaptive learning platforms 248
- Additional Protocol to the American Convention
on Human Rights in the Area of
Economic, Social and Cultural Rights
(San Salvador Protocol) 432
Adler, E.L. 290–91
administrative law 511–12, 513, 514–16
adversarial principle 296–97, 302–7
Aetna insurance company 232–33
Afghanistan 152–53, 327–28
Africa 332, 350–51
 data protection 154–55, 156–58, 159
 freedom of expression right 84–85, 88–89, 90
African Charter on Human and Peoples’ Rights
(ACHPR) 93, 105, 154–55, 159–61,
 295n.5, 386–87, 432
African Charter on the Rights and Welfare of the
Child (ACRWC) 154–55
African Commission on Human and People’s
Rights (ACommHPR) 154–55, 255, 433
African Cybersecurity and Personal Data
Protection Convention 154–55
African Declaration on Freedom of Expression
and Access to Information 154–55
African Union (AU) 213, 218
aggression detection 91
Agreement on the Trade-Related Aspects of
Intellectual Property Rights (TRIPS)
 Agreement 105–6, 108
AI advisory body 38
AI colonialism 351–52, 570
AI consultancy 38
AI decision-making limits 484–500
 causatives and declarative decisions
 distinction 490–95, 498–99
 consciousness 494–95
 doctors and diagnoses 488–91, 492–93, 494
 ethical issues 347, 486, 494, 498–99
 factual decisions 487–90, 491–92, 494, 496
 framework for decision-making by AI 498–99
 incompleteness of existing accounts 487–90

- AI decision-making limits (*cont.*)
 judges and judgments 488–90, 491–93,
 494, 497–98
 juries 497–98
 legal responsibility 496–97
 moral agents 494–95
 moral/ethical (normative) decisions 487–90,
 491–92, 493–94, 496–500
 why certain decisions should not be made by
 AI 485–87
- AI Global Surveillance Index 150–51
- AI lawyers 302–3, 485, 499
- AI register 40
- AI Sur report 157–58
- algorithm register 40, 304–5, 306–7
- algorithm-steered tunnel vision 69
- algorithmic redlining (housing) 359–60
- Alphabet 224–25
- alternative dispute resolution *see* online dispute
 resolution (ODR)
- Amazon 20n.9, 24, 257–58, 374–75, 376–77
 Alexa 253
 Comprehend Medical 391
 Echo 251–52
 fulfilment centres and ‘Amazon pace’ 373–74
 recruitment procedure 299–300
- American Convention on Human Rights
 (ACHR) 155–56, 159–61, 167, 265,
 295, 386–87
- American Declaration on the Rights and Duties
 of Man 105
- AnAbEL platform 365
- animal rights 477
- Annie MOORE software 321
- anonymity/anonymisation 9, 23, 24
 data analytics in justice system and
 healthcare 290
 fair trial rights 276–77
 food rights 352
 freedom of expression 79
 health rights 395–96
 human rights impact assessment (HRIA) 533
 LGBTQ+ rights 224–25
 public space, surveillance and right to be
 ignored 186–87, 188–91
see also pseudonymisation
- appeal and redress mechanisms 102
- Argentina 289
- arrest or detention 46–47, 49, 93
- Article 29 Data Protection Working Party (now
 European Data Protection Board 113–
 14, 138, 382, 536, 540, 549–50
- artificial general intelligence (AGI) 494–95
- Artificial General Super Intelligence (AGSI) 65
- Artificial Intelligence White Paper 406–7
- Asaro, P. 472–73
- Ascension 397
- Asia 84–85, 88–89, 90, 152–53, 157–61
- Asian Human Rights Charter 295n.5
- Asimov, I. 61
- assisted and augmented AI 19, 28–29
- assistive technologies 248, 250–52, 259
- Association of Southeast Asian Nations
 (ASEAN) 213, 218
- Human Rights Declaration 295n.5
- asylum rights 11, 74, 148, 286, 300, 311–26
 administrative judges 312
 asylum adjudicators 312
 country of origin information (COI) 319
 direct cash allowance 317–18
 electronic voucher programme 317
 food assistance 317–18
 forced migration 311–12
 forced migration management 322–24
 forced migration: positive and negative
 implications 312–22
 decision to flee and right to access territory
 of asylum 313–17
 durable solutions to end displacement
 cycle 321–22
 refugee status determination
 process 318–21
 right to file for asylum 317–18
 illegal interdiction measures 315
 immigration agents 312
 migrant deaths 315
 minimum procedural guarantees 312
non-refoulement (forced return) 314, 315, 316
 resettlement 321
 ‘smart border’ policies 314–15
 Syrian refugees in Jordan 317
see also asylum rights and screening
- asylum rights and screening 11, 327–39
 automated credibility tests 332–33
 exponential increase in applications 330–31
 international law 328–30
 manpower constraints 331
 migration management tools 331, 335
non-refoulement 328, 329–30
 policy proposal 337–38
 predictive analytics 331, 332, 333–34
 reasons for use of AI systems 330–31
 recommedatory automated decision-making
 (ADM) 333–34
 refugee status determination (RSD) 327–28,
 330–31, 335–36
 refugee status determination (RSD)
 officers 332–33

- Atkinson, J. 11–12
 audits 12, 36, 39, 133–34, 220, 277–78, 286–87,
 382–83, 422
 Augustine, Saint 472
 Austin, G. 105–6
 Australia 351–52
 data analytics 284–85
 Equal Opportunity Act 2010 205
 gender-based discrimination 213
 housing rights 364–65
 Human Rights Commission 226–27
 liberty and security rights 48–49, 51
 Modern Slavery Act 521
 privacy rights 130–31
 public space, surveillance and right to be
 ignored 179–80
 racial discrimination 204–5
 Regulatory Sandbox 563–64
 Target Compliance Framework and social
 security rights 300
 Australian Women Against Violence 240
 auto-identification 231–32
 automated decision-making (ADM) 10–11
 asylum rights 311–12, 316–17, 319–20, 322–
 23, 324, 325
 asylum rights and screening 331, 333–34
 aversion 410–11
 consumer protection rights 418–22
 disability rights 248, 250–51, 252–53
 freedom of expression 83–84
 housing rights 360, 361
 open-source 304–5
 privacy and personal data protection 125–26
 property rights and IPR 113
 smart cities and public law 514–15
 see also automated decision-making (ADM)
 and effective remedy rights
 automated decision-making (ADM) and effective
 remedy rights 294–307
 ability to process immense amounts of data 306
 fair hearing rights 295–98
 general characteristics of ADM 301–2
 generic AI applications 298–99
 legal framework 295–98
 obscenity 304–5
 positive and negative consequences 302–4
 automated facial recognition (AFR) *see* facial
 recognition technology (FRT)
 automatic gender recognition 147
 automatic operational restriction
 mechanism 34–35
 automation
 full 299, 300
 partial 299, 300
 autonomous (or automated) intelligence 29–30
 autonomous vehicles (AVs) 18, 253, 348–49,
 415–16, 423, 445
 AI decision-making limits 484–85, 498–99
 regulatory sandbox 557, 559, 560
 autonomous weapons system (AWS) 484–85,
 486–87, 499
 autonomy 10, 22, 281–82, 301–2, 396
 consumer protection rights 410–13, 422,
 423, 424
 Axinn, S. 486–87
 Ayyub, R. 238
 B-Tech Project 525
 Babic 400
 Baldus studies on death penalty 287–88
 Bangladesh 152–53
 Rohingya refugees 318
 Barocas, S. 223
Be My Eyes 251
 Belenguer, L. 195–96
 Belkhir, L. 430
 Benesty, M. 268–69, 286, 287
 Berne Convention 105–6, 108
 bias 10, 14
 AI decision-making limits 488–89
 asylum rights 312, 315–17, 319–20, 322, 324
 asylum rights and screening 335–37, 338–39
 audits 213
 biometrics 101, 103
 business and human rights (BHR) and private
 sector responsibility 526–27
 cognitive 410–11
 common risks and benefits 443–45, 446–
 47, 453
 consumer protection rights 416–17
 contextual 393
 data analytics in justice system and
 healthcare 286–87
 data preparation 19, 23
 deployment 33
 designer 335–36
 disability rights 248, 250–51, 252–53, 259
 effective remedy rights and ADM 303n.55
 fair trial rights 267–73, 277–78, 279–80
 freedom of expression 78
 gender-based discrimination 4, 210–11, 212–
 13, 214–15, 218–19, 220
 health rights 393, 402
 housing rights 369, 370
 irrational 484–86
 legacy 335–36
 LGBTQ+ rights 223, 234
 liberty and security rights 54–55, 56, 60

- bias (*cont.*)
 modelling 25
 nationality 334
 police 96–97
 privacy and personal data protection 126, 165
 privacy, personal data protection and FRT 147
 property rights and IPR 117–18
 racial/ethnic minorities 4, 195–96, 287
 representation issues 9–10
 sexual 287
 smart cities and public law 515
 transfer learning 16
 work rights and algorithmic
 management 376–77
see also discrimination
- Big Brother Watch* 100–1
- Big Data 12, 27–28, 350–51, 395–96, 418–19
 housing rights 360–61, 369
 privacy, data protection and FRT 139,
 142, 144
- Big Tech 528–29
- Bigman, Y.E. 487–88, 494n.48
- biometrics 7–8
 asylum rights 314–15, 317–18, 331, 332–33
 common risks and benefits 447–48, 452
 deployment 34
 disability rights 254, 257, 259
 indiscriminate, biased and opaque nature
 of 101, 103
 Kenya 298–99
 privacy and personal data protection 166–67
 privacy, personal data protection and
 FRT 136, 137, 139, 140, 142–44, 147–48
- public space, surveillance and right to be
 ignored 177, 183
- smart cities and public law 515
- templates 94
see also biometrics and surveillance under
 freedom of assembly and association;
 facial recognition technology (FRT);
 fingerprint scanning; iris-recognition;
 voice recognition
- Birch, T. 482
- 'black box' and opacity 12
 asylum rights 324, 337
 biometrics 101, 103
 common risks and benefits 445–46, 449–
 50, 456
 consumer protection rights 410–11, 412–13,
 420–21, 423
 effective remedy right and ADM 301–2,
 303, 304–5
 health rights 394–95, 402
 housing rights 360
- legal personality 464–65, 466
 liberty and security rights 60
 modelling 30
 privacy and personal data protection 127–
 28, 173
 privacy, personal data protection and FRT 142
 property rights and IPR 112, 113, 114–15
 systems restrictions 27–30
 work rights and algorithmic
 management 373, 378–79
- 'Black Lives Matter' 50
- blockchain 408–9
- Bloomer, P. 529
- Botero Arcila, B. 510
- boundary conditions 31–32
- Brandeis, L.D. 177
- Brazil 124–26, 130–31, 332
 General Law on Personal Data Protection 535
- Bridges v South Wales Police* and AFR Locate 58,
 180, 185, 186, 200–2, 203, 204–5, 206
- Bryson, J. 471
- Bufithis, G. 293
- Buolamwini, J. 54–55, 147
- business and human rights (BHR) and private
 sector responsibility 14, 517–30
 domestic and regional initiatives 521–24
 due diligence 517–18, 519–20, 522, 523–24,
 525–29, 530
 emerging framework 518–20
- Business Impact Analysis (BIA) 538n.48
- business understanding 17–18
- Bygrave, L.A. 548
- Canada 124–25, 127, 129, 130–31, 335,
 521, 554–55
- Department of Immigration, Refugees and
 Citizenship (IRCC) - advanced data
 analytics 320
- Immigration and Refugee Protection
 Regulations 300
- Known Traveller Digital Identity
 program 333–34
- 'Ontario Works' - Social Assistance
 Management System (SAMS) 300–1
- Čapek, K. 471
- CAPTCHA (Completely Automated Public
 Turing test to tell Computers and
 Humans Apart) tests 484n.4
- Carnegie Endowment for International
 Peace 150–51
- Castagnola, A. 289
- categorical approach 548, 549
- categorisation 138, 142, 147, 443–53
- Catt 99–100

- causative and declarative decisions
distinction 490–95, 498–99
- CCTV 48, 91, 142, 146, 156–58
- Ceron, M. 10
- Chakrabarty, I. 9
- chatbot therapists 251–52
- chatbots 418
- Chen, D. 271
- child, rights of 147–49
- Chile
Marcha del Millón protests (2019) 93
- National Board of Scholarships and School Aid FRT for school meals distribution 298–99
- chilling effect 7–8, 136, 146, 156–57, 183, 379, 380, 420–21
- freedom of assembly: biometrics and surveillance 92, 98–100, 101
- China 48–49, 159, 188n.84, 196, 333–34, 364–65
Beijing Olympics 180n.17
Communist Party 156–57
public space, surveillance and right to be ignored 179–80, 181, 184
- re-education camps 68, 156–57
- Uyghur ethnic minority community
surveillance 68, 71–72, 147, 156–57, 454
- WeChat 226
- Church of Scientology 66–67
- civil law 511–12, 515–16
- civil and political rights 5, 7–8, 9, 228–29, 352, 451–52, 517, 564–65
- civil society organisations 528–29
- CLAUDETTE 407–8
- clickbait 34
- Clinton, H. 238
- closed systems 24
- co-regulatory schemes 86
- 'Coded Bias' documentary 551–52
- codes of conduct 258–59
- codification 395–96
- Coeckelbergh, M. 479
- Cofone, I. 8–9
- Cogito 374
- CogniCity software 363
- cognitive computing 276–77, 280
- cognitive correction 287
- Cohen, G. 400
- coherence-focused rules 274
- collective action engagement 527–28
- collective bargaining 11–12, 372, 383–85
- collective solidarity rights 5, 12
- Collins, P. 11–12
- colonial technologies of racialised governance 11
- Commission on the Status of Women (CSW) 216–17
- Commission on Transnational Corporations 518
- Committee on Economic Social and Cultural Rights (CESCR) 357
- Committee on the Elimination of Racial Discrimination (CERD) 199
- commodification of physical features 143–44
- common risks and benefits 441–57
categorisation 443–53
common benefits 451–52
epistemic concerns 453, 455
functional risks 443–44, 446–51, 453, 455–56
legal safeguards for substantive legal protection 448–51
substantive dimension 446–48
- negative dimension 451–52
- normative concerns 453, 455–56
- political legal solutions 443–44
- positive dimension 451–52, 453
- public interest and private ramifications 453–56
- structural risks 443–46, 453, 455, 456
- technical solutions 443–44
- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)
software and *Loomis* case 52–53, 58, 113, 197, 271–72, 275–77, 278–79, 287–88
- Compton, C. 11
- confidentiality 39–40
commercial information 110
commercial information *see also* trade secrets
- privacy and personal data protection 131–33
privacy, political participation and FRT 154–55
- property rights and IPR 117–18
- public space, surveillance and right to be ignored 187–89, 190–91
see also anonymity/anonymisation; sensitive data
- conflict management 409
- conformity assessment 259
- conscientious objection, right to 289–90
- consciousness 13–14, 462–66, 470, 480–81, 494–95
- consent
asylum rights 322–23, 337
common risks and benefits 449–50, 456
consumer protection rights 418–20
data analytics in justice system and healthcare 290
- data preparation 24
- disability rights 252

- consent (*cont.*)
 express 325
 free 173, 317–18
 health rights 396–97
 informed 168, 172–73, 317–18, 394, 411–12
 non-consensual dissemination of deepfake images 237–38, 241, 245, 247
 obligatory 246
 privacy and personal data protection 129, 171–73
 privacy, personal data protection and FRT 140–43, 147–48
 regulatory sandbox 560, 564–65, 566
 specific 173
 withdrawal 173
 consequentialist approach 548, 549
 consumer manipulation 12
 consumer protection rights 12, 73, 74, 111–12, 405–24, 446–47, 452, 517, 564–65
 abusive market behaviour 405–6
 ad-blockers 407–8
 advertising, targeted and personalised 411–12, 419–20, 421–22
 anti-phishing 407–8
 anti-spam 407–8
 approaches and challenges to improve status quo 419–24
 autonomy of consumers 422, 423, 424
 burden of proof 414
 consumer protection as human right 405–7
 counterfeit products 408–9
 design defects 415–16
 e-commerce 409, 412, 413–14, 423–24
 immaterial harm 416
 justice, access to 409
 loss of chance 416
 negative aspects of AI 410–19
 discrimination 12, 416–19
 personal autonomy 410–12
 product safety and product liability 412–16
 non-harmonised products 413–14
 online games 415–16
 online platforms 422–23
 online shopping 407–8, 409, 411–12
 pain and suffering 416
 passenger right claims 409
 positive aspects of AI 407–9
 price comparison websites 407–8
 price discrimination 418–19
 price personalisation 421–22
 product safety and product quality 408–9
 terms of service 407–8
 traceability of products and product recalls 408–9
 unfair contractual clauses 407–8
 content curation 80–81, 82–83, 88, 89–90, 398–99
 content moderation 7–8, 76, 82–83, 84–85, 88–90, 111–12
 hard 82–84
 hybrid 83
 soft 82–83
 content over-removal 89–90
 content prioritisation, faulty or biased 7–8
 contestability, limited 448–50, 456
 context of AI 542–43
 contextual factors 31–32, 445–46, 455–56
 contract, performance of 125
 control 39
 guarantees 35
 Convention 108+ 124–25, 137, 139–40, 147–48, 535, 561
 Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) 207–8, 212–13, 220–21
 Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) Committee 212, 216–17
 Convention on the Rights of the Child 148
 Convention on the Rights of Persons with Disabilities (CRPD) 364–65
 copyright protection 84, 106–8, 109–10
 CoreLogic 360
 corporate irresponsibility and impunity 14, 520
 corporate responsibility 436, 519, 528–29
 Corti 391
 Cottier, B. 154–55
 Council of Europe (CoE) 86, 207–8, 213, 214, 218–19, 228–29, 255
 Commission for the Efficiency of Justice (CEPEJ) 277–79
 Commission for the Efficiency of Justice (CEPEJ), Ethical Charter 278, 280
 Committee of Ministers 56, 86
 counterfeit products 408–9
 court and judge analytics *see under* data analytics in justice system and healthcare
 Covid-19 pandemic 32–33, 145–46, 197, 343–44, 372, 412
 Crawford, Judge 199
 Crawford, K. 95–96
 credibility 29, 39, 332–33, 563
 creditworthiness 542
 criminal activity management 50–51, 53
 criminal justice 7, 196, 197
 see also law enforcement and criminal justice
 CrimSAFE 360
 critical infrastructures 35
 Cross-Industry Standard Process for Data Mining (CRISP-DM) 17n.2

- cruel punishment and treatment, rights
against 475–76
- cryptocurrency mining 430
- currency of rights 476
- customary international law 228–29
- customisation of content 81, 89
- cybercrime/cyberattacks 69, 412–13, 420–21
see also hacking
- cybersecurity 12, 32–33, 36, 399, 402, 429, 431, 444, 445, 456
- 'Daily Me' 81
- damage prevention and damage control 70
- Danfoss* 112–13n.55
- dangerous content *see* illegal and harmful content
- data absorption phase 32–33
- data altruism 401
- data analytics 10, 350–51
- data analytics in justice system and healthcare 281–93
- court and judge analytics 10, 281–82, 283–89
French ban on judge analytics 286–87, 292
French ban and personal data protection laws 287–89
impartiality and autonomy of judges 281
mainstreamed judicial analytics 285
- descriptive analytics 283–84, 291
- healthcare and physician analytics 281–83, 289–93
- predictive analytics 283–84, 286, 291
prescriptive analytics 283–84, 291
- data breaches 131, 140
- data cleansing 18–19
- data collection 23, 173, 179, 318, 352, 425
consumer protection rights 408–9, 411–12
LGBTQ+ rights 224–26, 233–34
privacy rights 136, 148
work rights 371–72, 374
- data controllers
human rights impact assessment (HRIA) 538–40, 541
- privacy and personal data protection 170, 171, 172, 174
- privacy, personal data protection and FRT 141–42
- privacy rights 124–26, 127, 128–31, 133, 134–35
- data dependency 412–13
- data diversity 102
- data extraction 416–17
- Data Governance Act 401
- data governance systems 381–82
- data maximisation 130
- data minimisation 128, 129–30, 132, 134–35, 306
- data mining farms 45
- data preparation 17, 18–24, 33, 39–40, 446–47
biased data 23
discrimination prohibition 19–21
privacy rights 23–24
- data processing during deployment 32–33
- data processors and privacy rights 124–25, 126, 128, 129–31, 142
- data protection 4–5, 8–9
Africa 154–55
freedom of assembly: biometrics and surveillance 102
freedom of expression 78–79, 89
public space, surveillance and right to be ignored 178, 181
religious freedom 62–63
see also personal data protection; privacy and personal data protection
- Data Protection Act 382
- Data Protection Impact Assessments (DPIAs) 14–15, 133–34, 382–83
see also under human rights impact assessment (HRIA)
- Data Protection Officer (DPO) 133, 540–41
- data quality 102, 398–99
- data retention 102
- data sharing 303–4, 306–7, 317, 352
- data solidarity 401
- data storage periods 102
- data theft 445
- data transformation 18–19
- datafication 291
of facial features 144
of gang policing 96–97
- De Gregorio, G. 7–8
- de Heer, S. 10–11
- De Hert, P. 14–15
- de-anonymisation 136
- debiasing procedures 59
- decision-making 13, 14
see also AI decision-making limits; automated decision-making (ADM)
- decision support systems (DSS) 269–70, 277–78, 279–80
- decision trees 25–26
- deep AI 65–66
- deep learning (DL) 224–25, 239, 420–21, 444–45
- deepfake technologies 9–10, 34, 77–78, 224–25
see also women's rights and deepfake pornography
- DeepMind 397
- DeepNude app 240

- defamation offences 241–42
 Deliveroo 378
 democratic right 7–8
 deployment of AI 31–36
 abuse of AI technology 36
 data processing during operation 32–33
 malicious use of AI technology 34–35
 monitoring 36
 not matching target operating environment 31–32
 descriptive analytics 283–84, 291
 design phase 5, 7, 71, 537–38
 deterrence effect *see* chilling effect
 development of AI 18, 30, 517, 521, 527, 537–38
 Dewey, J. 458n.1
 Dialect Identification Assistance System 332–33
 differentiation of rights 478–82
 digital authoritarianism 9, 156–58
 digital phenotyping 389–90
 digital self-determination 5
Dima v Romania 107–8
Dinerstein v Google 397–98
 disability rights 10, 248–61
 accessibility 259
 activity limitations 249–50
 assistive technologies 248, 250–52, 259
 drawbacks of AI technologies 252–54
 medical models of disability 250
 mental functioning limitations 249–50
 new regulatory solutions and their impacts 258–59
 participation limitations 249–50
 representation 260
 social and material contexts 250
 social models of disability 250
 work rights and algorithmic management 376–77, 382
see also United Nations Convention on the Rights of Persons with Disabilities (CRPD)
 disadvantaged groups *see* ethnic minorities, minorities and disadvantaged groups
 disaster risk reduction 11, 362–64, 369–70, 429
 disclosure 40, 114–15, 116–17, 259, 521, 522–23
 discrepancies 36
 see also errors; inaccuracy
 discrimination 8, 10, 14
 asylum rights 316–17
 business and human rights (BHR) and private sector responsibility 526–27
 common risks and benefits 446–47, 453, 456
 consumer protection rights 12, 416–19, 420–21, 423–24
 criminal courts 460
 direct 19, 416–17
 disability rights 250–51, 261
 fair trial rights 271
 freedom of expression 78, 84, 85, 88–89
 housing rights 13P8, 360, 361–62, 369
 indirect 19, 27, 41, 417, 418–19, 424
 legal personality 469–70
 LGBTQ+ rights 234
 liberty and security rights 45, 59
 price 418–19
 privacy, personal data protection and FRT 139–40, 147
 property rights and IPR 112–13
 proxies 210
 religious freedom right 68, 69, 71–72
 smart cities and public law 503–4, 508–9
 unjustified 33, 34
 unlawful 443–44
 work rights and algorithmic management 375–77, 378
see also gender-based discrimination; non-discrimination; racial discrimination
 Displacement Tracking Matrix 332
 dissent *see* privacy, political participation and FRT
 distorted data 19
 distorted images and videos 34
 distributed ledger technology (DLT) 349–50
 doctors and diagnoses 488–91, 492–93, 494
 domestic law 58, 328–29, 431, 433–34, 520, 524
 DoNotPay ‘robot lawyer’ 302–3
 Drchal, J. 7, 14–15
 drones 170–71, 179–81, 184, 311–12, 315, 348–49, 447–48
 killer 34, 499
 dual-use items 72
 due diligence 9, 14
 business and human rights (BHR) and private sector responsibility 517–18, 519–20, 522, 523–24, 525–29, 530
 common risks and benefits 454–55, 456–57
 consumer protection rights 423–24
 disability rights 258
 discrimination prevention 202–5, 206
 effective remedy right and ADM 305
 healthy environment right 436–37
 housing rights 360–61
 human rights impact assessment (HRIA) 534–35, 544–45
 LGBTQ+ rights 230
 due process rights 26, 40, 113, 379–80
 asylum rights 316–17, 319, 320, 329
 Dunn, P. 7–8
 Dworkin, R. 274–75, 276–77, 280

- E-meter 66–67, 72–74
 earth-friendly usage of AI 13
 eBay Resolution Centre 409
 Ebert, I. 14
 echo chambers 81
 Economic Community of West African States (ECOWAS) - Supplementary Act on Personal Data Protection 154–55
 economic rights *see* economic, social and cultural rights
 Economic and Social Commission for Asia and the Pacific 255
 economic, social and cultural rights 5, 11, 74, 105–6, 228–29, 370, 386–87
 education, right to 113, 257–58, 327–28, 570
 effective remedy right 10–11, 12, 86, 112, 520
see also automated decision-making (ADM)
 and effective remedy rights
 efficiency of output 29
 Electronic Health Records (EHR) 290–91, 388
 Ellie (virtual therapist) 389
 Elmeligi, A. 430
 emotion recognition 68, 95–96, 254, 379
 employment rights *see* work rights
 encryption 79, 102, 408–9
 environmental rights 12
 epidemic prediction 390
 epistemic factors 453, 455, 480–81
 equality 4, 9, 10
 accommodating 256, 257–58, 260
 data analytics in justice system and healthcare 293
 disability rights 248, 256, 260
 fair trial rights 268–69
 freedom of expression 78–79
 inclusive 256
 LGBTQ+ rights 223, 226, 227–29
 participative 256, 257, 260
 privacy, personal data protection and FRT 147, 149
 procedural 296–97
 racial discrimination 10, 202–3
 recognition 256, 257, 260
 redistributive 256, 260
 religious freedom, right to 71–72
 smart cities and public law 503–4
 work rights and algorithmic management 371–72, 376, 377–78, 380
 equality of arms principle 266, 296–97, 302–7, 503–4
 erasure/deletion 130–31
 erroneous design 444–45, 456
 errors 7, 36
 asylum rights 316–17, 335–36, 338
 freedom of expression 83–84, 88–90
 health rights 393–94, 402
 Esposito, R. 482, 548
 ethical/moral issues 292–93, 401
 AI decision-making limits 485, 486, 489–90, 494, 498–99
 asylum rights and screening 337, 338–39
 human rights impact assessment (HRIA) 547–48
 moral agents 494–95
 moral status 479
 moral subjects 13–14
 moral/ethical (normative) decisions 486, 487–90, 491–92, 493–94, 496–500
see also natural (moral) rights
 ethnic minorities, minorities and disadvantaged groups 4, 25, 56, 84–85, 89–90, 95, 96–97, 147, 169, 360, 392, 502–3, 512, 526–27
see also marginalised groups
 Eubanks, V. 190
 Eurodac 148
 European Blind Union 259
 European Charter of Patients' Rights 290
 European Commission 332, 532
 Proposal for a Directive on Corporate Sustainability Due Diligence 523–24
 European Commission on Human Rights (ECommHR) 66–67, 73, 169
 European Convention on Human Rights (ECHR) 58, 63–64
 data analytics in justice system and healthcare 284
 effective remedy right and ADM 295
 fair trial rights 265
 freedom of assembly: biometrics and surveillance 93, 99–101
 freedom of expression 76, 77
 gender-based discrimination 213
 health rights 386–87
 healthy environment right 435–36
 human rights impact assessment (HRIA) 532
 privacy and personal data protection 121–22
 privacy protection 167, 169–70
 property rights and IPR 105, 107–8, 111–12
 regulatory sandbox 552–53, 557, 559, 560, 562, 565
 work rights and algorithmic management 377–78, 379–80, 382

- European Court of Human Rights (ECtHR) 26, 86, 99, 103, 107–8, 122–23, 146, 169
 common risks and benefits 447–48
 data analytics in justice system and healthcare 284
 effective remedy right and ADM 295
 fair trial rights 266–67, 269–70
 freedom of assembly: biometrics and surveillance 92
 health rights 386–87
 HUDOC database 265
 Open Data project 277–78
 privacy, personal data protection and FRT 143
 public space, surveillance and right to be ignored 186
 regulatory sandbox 552–53, 559, 563
 work rights and algorithmic management 379
- European Data Protection Supervisor (EDPS) 143–44, 396–97
- European Disability Forum 259
- European Ethical Charter on the use of Artificial Intelligence in Judicial Systems and their Environment (2018) 277–78
- European Health Data Space (EHDS) proposal 401
- European High Level Expert Group on Artificial Intelligence 535–36n.32
- European Medicines Authority (EMA) 395
- European Parliament (EP) 227–28, 532–33
- European Patent Convention (EPC) 108–9
- European Social Charter 110–11
- European Social Committee 358–59
- European Strategy for Data 401
- European Union
 asylum rights 323
 consumer protection rights 407, 412
 data analytics in justice system and healthcare 286
 fair trial rights 271, 278–79
 food rights 350–51
 freedom of expression 87–89
 gender-based discrimination 213
 housing rights 364–65
iBorderCtrl 315–16
 legal personality 461
 predictive policing 51
 privacy and personal data protection 130–31
 property rights and IPR 108, 112–13
 public space, surveillance and right to be ignored 181, 185
 religious freedom, right to 71–72
 smart cities and public law 510
- European Union Agency for Fundamental Rights 55
- European Union Agency for the Operational Management of Large Scale IT systems (eu-LISA) 148
- European Union Artificial Intelligence Act (proposed) 56–57, 72, 117–18, 184, 218–19, 512, 529–30
 asylum rights 323
 consumer protection rights 406–7
 disability rights 255, 258–59, 261
 human rights impact assessment (HRIA) 532–33, 536, 537–38, 541–42, 545
 registration 317
 regulatory sandbox 552, 554–55, 556–57, 558, 559n.35, 560–61, 566
 work rights and algorithmic management 381–83
- European Union Audiovisual Media Services Directive 70–71
- European Union Border and Coast Guard Agency (FRONTEX) 315
- European Union Charter of Fundamental Rights (CFR) 70–71, 121–23, 167, 295, 382, 386–87, 544, 557n.29
 consumer protection rights 405–7
 property rights and IPR 105, 110–12
- European Union Data Protection Authority (DPA) 97, 141–42, 145–46, 167, 170, 171, 175–76
- European Union Data Protection Directive 122–23
- European Union Data Protection Supervisor (EDPS) 546–47
- European Union of the Deaf 259
- European Union Digital Markets Act (DMA) 421–23, 523n.23
- European Union Digital Services Act (DSA) 88, 90, 421–22, 523, 529–30
- European Union General Data Protection Regulation (GDPR) 104, 113–14, 529–30
 asylum rights 322–23
 consumer protection rights 420n.116, 422
 data analytics in justice system and healthcare 282–83, 288
 effective remedy rights and ADM 306–7
 privacy and personal data protection 122–26, 127, 129, 133, 166, 168, 171–72
 privacy, personal data protection and FRT 137, 139–40, 141–43, 147–48
 regulatory sandbox 566
 work rights and algorithmic management 382–83
see also under human rights impact assessment (HRIA)

- European Union General Product Safety
Directive (GPSD) 413–14, 415
- European Union Privacy Commissioners 167
- European Union Product Liability Directive (PLD) 414, 415
- European Union Proposal for a Regulation on General Product Safety 413–14, 415
- European Union Trade Secrets Directive (EUTSD) 110, 115–18
- European University Institute 407–8
- evaluation 30
metrics 25
- Evgeniou 400
- ex post* information and testing mechanism to third parties 39–40
- ex post* review methods based on a counterfactual 27, 30, 37
- exception clause 560, 562
- exceptional cases 101
- exceptions 118, 140
- exclusion 14, 508–11, 515
- exclusive rights 116–17
- executions and firing squads 497–98
- explainability 28–29, 30
common risks and benefits 445–46
consumer protection rights 420–22
fair trial rights 280
human rights impact assessment (HRIA) 537–38
- human rights risk assessment 37, 38–39
- privacy and personal data protection 134–35
- property rights and IPR 111–12, 114–15
see also unexplainability
- explanation, right to 113–14
- eye-tracking 248, 251
- Face Recognition Vendor Test on Demographic Effects (National Institute of Standards and Technology (NIST) 95
- Facebook 84–85, 246n.79, 292–93, 389
- facial recognition technology (FRT) 8–9
asylum rights 311–12, 314–16
asylum rights and screening 330, 332–34, 335–36, 337
- business and human rights (BHR) and private sector responsibility 526–27
- Chile 298–99
- common risks and benefits 447–48, 452
- data preparation 21–22, 23
- disability rights 253
- freedom of assembly: biometrics and surveillance 91, 95, 97, 98–99
- law enforcement and criminal justice 48–49
- LGBTQ+ rights 226–27
- liberty and security rights 53–55, 57–58
- privacy and personal data protection 124, 126, 128, 132–33, 134
- property rights and IPR 113
- public space, surveillance and right to be ignored 179–81, 184
- racial discrimination 196, 200–2
- work rights and algorithmic management 376–77, 382
- see also* privacy, personal data protection and FRT; privacy, political participation and FRT
- facial signatures 48
- facilitate, duty to 74, 346–47
- factual decisions 487–90, 491–92, 494, 496
- Fadeyeva v Russia* 435
- failsafe mechanisms 39
- fair hearing right 295–98, 302–4, 306–7
- fair trial rights 10, 26–27, 30, 265–80, 503–4
accusation made promptly and in understandable language 266
adequate time and resources to mount defence 266
anonymisation, automatic, of judgements 276–77
bench of judges rather than single judge 269
bias 272–73, 277–78, 279–80
AI judges 271–73
human judges and inefficient courts 267–70
- citation networks 276–77
- common risks and benefits of AI 445–46, 448–49, 450–51
- consistency and coherence 279–80
- decision-making limits of AI 485, 499–500
- effective remedy rights 295–96, 306–7
- equality of arms 266
- free legal assistance if required 266
- Hercules (super-human judge): division of labour between humans and computer 273–77, 280
- objective ignorance 272–73
- obligation to give reasons 269
- precedents 276–77
- presumed innocent until proven guilty 266
- property rights and IPR 112, 113
- regulation attempts 277–79
- securing consistent and impartial adjudication 266–67, 268–69
- securing justice within a reasonable time 266–67
- statistics and machine learning prohibition for judges 268–69
- transparency 39
- witnesses and interpreters 266

- fair wages right 353
 fairness 10, 12
 asylum rights 319–20
 consumer protection rights 424
 disability rights 252, 254
 LGBTQ+ rights 233–34
 liberty and security rights 59
 privacy and personal data protection 125–26,
 127–28, 131
 procedural 4–5, 320
 fake AI techniques 72–73
 see also deepfake technologies
 fall back procedures 102
 false alarms 429
 false information 69
 false negatives 338
 false positives 83–84, 89–90, 147, 338, 526–27
 Favalli, S. 10
 Feldstein, S. 156–57
 filter bubbles 81
 filters 78, 84, 233–34, 245–46
 fingerprint scanning 148, 384
 Finland 304–5
 Fintech companies 554, 557, 560–61, 563–64
 first generation rights 6, 11–12, 453
 Floridi, L. 181–82, 471, 479, 494–95
 food rights 11, 343–54, 452
 accessibility, economic and physical 346
 conflicting rights 352–53
 food supply chains (FSCs) 344, 345–46, 347–
 48, 349–51
 greenhouses and vertical and hydroponic
 farming 349
 inequality 350–52
 meaning and scope of right to adequate
 food 344–48
 progressive realisation 346–47
 smart farming 344, 348–53
 Food Systems Summit 343–44
 forced migration 11
 see also under asylum rights
 foresight techniques 546, 548
 Foss-Solbrekk, K. 484n.2
 Foucault, M. 231
 fourth generation rights 5, 462n.16
 fragmented data 19
 France 268–69, 278, 287–89
 CNIL Sandbox Initiative for Health
 Data 554–55
 judge analytics ban 286–87, 292
 Loi de Vigilance 522–23
 ‘Obligation to leave the French territory’
 (OQTF) expulsion measures 286
 Tribunal administratif de Marseille 141–42
 fraud 72–73, 272, 317, 332–34, 379–80
 see also social welfare fraud enforcement
 Freedman, S. 204–5
 freedom of assembly and association 7–8,
 11–12, 140, 149, 156–57, 183, 226, 447–
 48, 452
 biometrics and surveillance 91–103
 chilling effects 92, 98–100, 101
 European Convention on Human Rights
 (ECHR) 93, 99–101
 faulty biometrics 98
 government abuse 101
 inaccuracies 94–96
 indiscriminate, biased and opaque
 nature 101, 103
 indiscriminate or non-targeted
 surveillance 100–1
 positive obligations 93–94, 103
 promises 93–94
 safeguards 101–3
 soft biometrics 95–96
 underlying practices 96–99
 privacy, personal data protection and FRT 146
 work rights and algorithmic
 management 377–78, 380
 freedom of expression and information 7–8,
 11–12, 76–90, 156–57, 183, 226, 234, 257,
 298, 299–300
 active dimension 77, 86, 87–88
 passive dimension 77, 86, 87–88
 private governance 81–85
 property rights and IPR 106–7, 110–12, 116–
 17, 118
 regulatory sandbox 560
 state and business obligations 85–89
 work rights and algorithmic
 management 379, 380
 freedom of thought 544
 freedom to conduct a business 110–11
 Friends of the Earth 522–23
 fulfil, duty to 7, 74, 346–47
 functional risks 13, 443–44, 446–51, 453,
 455–56
 legal safeguards for substantive legal
 protection 448–51
 substantive dimension 446–48
 function creep 102, 156–58, 448
 fundamental rights
 IPR 110–15
 regulatory sandbox 556–58
 Fundamental Rights Agency 147–48
 gait recognition technology 298
 Gaius 482

- Gamper, F. 14
gaze detection 253
GCHQ ‘Tempora’ 447–48
Gebhard 145–46
Gellers, J. 477, 478
gender-based discrimination 207–21, 517–18
 gender data gap 209–10
 human rights law (HRL) 215
 hybrid approach 217–21
 human rights law (HRL) framework 208–9, 217–18
 regulation, elements for 218–20
negative effects 208–10
negative effects mitigation 211–20
 existing framework and limitations 212–17
 hard law 212–13, 215
 soft law 211–12, 213–15, 218–19
positive effects 208–9, 211–12
United Nations institutions and
 approaches 207–8, 210, 211–13, 215–17, 218–19, 220–21
women’s rights 9–10
 women’s rights and deepfake
 pornography 236–39, 241, 247
generalisation test 467–68
generative adversarial networks 239
Geopolitica (formerly PredPol) 51
geolocation restrictions 34–35
Gerke, S. 400
Germany 48–49, 168, 514–15
 Federal Office for Migration and Refugees (BAMF) 332–33
 Holocaust 318
 Supply Chain Diligence Act 523
‘gig economy’ 372
Giggle app 226–27
Gillespie, T. 82–83
GLAAD (Gay and Lesbian Alliance Against Defamation) 224–25
‘golden data sets’ 400
Golunova, V. 7
good administration principles 504–5, 512
Google 195–96, 224, 377–78, 397–98, 509
 photo recognition classification 485–86
Gordon, J.-S. 480
Graaf M. de 475
graphical models 25–26
Gravett, W.H. 156–57
Greece 97, 315–16
Greenleaf, G. 152–55, 157–58
Grey, K. 487–88, 494n.48
grievance mechanisms 520
 see also online dispute resolution (ODR)
Grimmelmann, J. 82–83
Gromova, E. 15
Gunkel, D. 13–14
habeas data 5, 155–56
hacking 36, 132, 317, 399, 429
Hamburg G20 protests (2017) 93, 95–96
Hanson Robotics ‘Sophia’ humanoid robot 473–74
Harari 481
hard law 212–13, 215
harmful content *see* illegal and harmful content
Harpur, P. 202–3, 206
Harris, D.J. 561
Hart, R.D. 473–74
hate crime 224–25
hate speech or disinformation 84–85, 88–89, 252, 517–18, 528
hateful incitement dissemination 69–70
hatred and/or violence incitement 70–71
health bots 71, 74
health rights 12, 113, 380, 386–402, 445, 449–50, 452
 challenges 392–99
 inaccuracy and errors 393–94
 privacy 395–99
 unexplainability and opacity 394–95
 clinical care and healthcare management 391
 digital phenotyping 389–90
 Electronic Health Records (EHR) 290–91, 388
 entitlements 387
 epidemic prediction 390
 ethical issues 401
 freedoms 387
 inclusive right 387
 individual right 387
 language understanding applications 391
 legal issues 401
 mapping health-related applications 387–92
 mental health 389
 policy and regulatory proposals 399–402
 precision medicine 388
 primary uses of health data 396
 public health 390
 public and private interest 396–97
 secondary uses of health data 396–97, 398
 social dimension 387, 398
 technical issues 401
 translational research 390–91
 update problem 395
 volume, velocity and variety 398–99
health and safety 226–27
healthcare analytics 281–83, 291–93

- healthcare sector 4, 10, 12, 17, 23, 27–28, 29, 105–6, 197, 327–28, 445
 disability rights 251–53
 health rights 387, 391, 393–94, 396–97
 HIV pre-exposure treatments and outcomes 224
 LGBTQ+ rights 223, 231
 regulatory sandbox 554–55, 566
see also data analytics in justice system and healthcare
- HealthMap 390
 healthy environment right 425–38, 444–45
 AI as pollutant/energy intensiveness/carbon footprint 429–31, 434–53
 climate informatics 427
 corporate actors 436
 current challenges 434–37
 defining right to environment 431–34
 earth-friendly AI 426–28, 437
 lack of international right 435–36
 misuse 429, 431
 negative consequences 429–31
 quantification problem 434–35
 unsustainability of AI infrastructure 429–30
 urban dashboards 428
 heatmaps 377–78
 Heine, K. 13–14
 Helfer, L. 105–6
 Helsinki AI Register 40
 high-risk AI technology *see* human rights-critical AI
 Hildebrandt, M. 507–8
 Hindriks, F. 475
 Hindriks, K. 475
 HireVue 372
 Hobbes, T. 468–69
 Hohfeld, W. 472, 476–77
 Hohmann, J. 11
 holistic approach 6, 16, 41, 132, 258, 338–39, 428, 457
 Hong Kong protests (2019) 93, 156–57
 horizontal effect 87–88
 housing rights 11, 355–70, 446–48, 452
 algorithmic redlining 359–60
 automated evictions 359–60, 361
 automated tenant screening services 359–60
 cognitive disabilities and dementia 365–66
 defining right in international law 356–59
 disaster risk reduction 362–64, 369–70
 discrimination 197
 emerging violations 359–62
 older persons/aging population 364–65
 persons with disabilities (PWD) 364–65
 remaining in place 364–67
 sanitation 367–69
 tenant screening services 360–61
 Hsin, L. 14
 Huawei 156–57
 Hubbard, L.R. 66
 human dignity 141–42, 143–44, 149, 257–58, 282, 394, 406–7, 410, 486–87, 488–89
 human fallibility 10
 human responsibility *see* robot rights/human responsibility
 human rights impact assessments (HRIAs) 6, 7, 13, 14–15, 258, 259, 531–50
 asylum rights 325
 baselines approach requiring further refining 549–50
 common risks and benefits 454–55, 456–57
 consumer protection rights 423–24
 Data Protection Impact Assessment (DPIA) 532–38, 549–50
 Data Protection Impact Assessment (DPIA) in GDPR 538–41
 effective remedy right and ADM 305
 EU Artificial Intelligence Act (proposed) 532–33, 536, 537–38, 541–42, 545
 from DPIA to HRIA-AI 541–49
 assessment of risks to rights 548–49
 mitigation measures 549
 necessity and proportionality 546–48
 pre-assessment 541–42
 systematic description of AI system 542–46
 GDPR 532–33, 535–37, 541, 542, 549–50
 liberty and security rights 59
 human rights law (HRL)
 common risks and benefits 441–42
 fair trial rights 274
 framework for algorithmic discrimination 208–9, 217–18
 freedom of expression 86–87
 gender-based discrimination 215
 housing rights 356, 360–61, 370
 LGBTQ+ rights 230
 regulatory sandbox 558
 see also international human rights law (IHRL)
 human rights risk assessment 37–39, 41, 86
 Human Rights Watch (HRW) 318
 human rights-critical AI 28, 30, 37, 38–39, 41
 human trafficking, forced labour and slavery 521
 human-centric approach 260–61
 Hume, D. 466–67, 468
 Hungary 315–16
 Aliens Policing Authority 332–33

- IBM 224
 Cúram Social Program Management 300–1
 Watson 395n.95
 identification (one-to-many) function 138,
 145–46, 147
 identity theft 317
 IDx-DR 387–88
 ignored, right to be *see* public space, surveillance
 and right to be ignored
 illegal activity 116–18
 illegal and harmful content 70–71, 77–78, 82–
 83, 84, 85, 88–89
 illegal loggers 429
 illegitimate use of AI 448, 456
 illiteracy, technical 301–2
 impact assessments 218–19, 385, 422, 563–64
 see also Data Protection Impact Assessments
 (DPIAs); human rights impact
 assessments (HRIAs)
 impartiality 286
 inaccuracy 12
 asylum rights and screening 336–37, 338–39
 effective remedy right 302, 304
 gender and ethnicity 145–46
 health rights 393–94, 395, 402
 privacy, personal data protection and
 FRT 145–46
 see also errors
 Inaytaullah, S. 473
 incomplete data 19
 incomplete input data 16, 302, 304
 India 48–49, 129, 136, 152–53, 333–34
 Ahmedabad smart city 508–9
 Delhi High Court and police use of FRT 97
 protests (2021) 93
 Supreme Court homosexuality
 decriminalisation 224–25
 individualised control versus government
 regulation 9
 indivisibility of human rights 6
 Indonesia 152–53, 226
 Jakarta 362–64
 Peta Jakarta 363
 inefficiency 269–70, 350–51
 infobesity 449–50
 information access rights 7–8, 40, 113–14, 115–
 18, 127, 128, 257, 281, 290–91, 560–61
 information asymmetry 405, 411–12, 419–21
 information quality 80–81
 information requirements 213
 informational duties 74
 informational self-determination 168
 InfoSoc Directive 109–10
 inner freedom (*forum internum*) 63–64
 input analysis 17–18, 37
 input data 16, 20, 21–22, 23, 24, 27, 28, 29–30,
 32–33, 38, 39–40, 319–20
 asylum rights 319–20
 effective remedy rights 301–2
 fair trial rights 272
 housing rights 366
 input layer 127
 Instagram 292–93
 Instagram, Community Guidelines 299–300
 integration theorem 467
 integrity 131–33, 185, 464, 465
 intellectual property rights 8, 39–40, 301–2,
 303, 304–5, 306–7, 420–21, 449–51,
 470
 see also property rights and intellectual
 property rights
 Inter-American Convention on Human
 Rights 93
 Inter-American Court of Human Rights
 (IACtHR) 155–56, 231–32, 269
 interconnected devices 415
 interdependence of human rights 6
 interest theory 477–78
 intermediary services 82
 see also third parties
 International Association for Impact
 Assessment 538n.45
 International Bill of Rights 520
 International Convention on the Elimination
 of All Forms of Racial Discrimination
 (ICERD) 198–99, 200, 202, 203–4,
 206
 International Court of Justice (ICJ) 199
 International Covenant on Civil and Political
 Rights (ICCPR) 45, 63–64, 76, 93,
 121–22, 159, 167
 AI decision-making limits 485
 asylum rights and screening 328–30
 effective remedy right and ADM 295
 fair trial rights 265
 robot rights/human responsibility 475
 work rights and algorithmic
 management 377–78
 International Covenant on Economic, Social and
 Cultural Rights (ICESCR)
 food rights 343–44, 345, 346–47, 352,
 353–54
 health rights 386–87
 housing rights 356–57, 359
 property rights and IPR 105, 106–7
 robot rights/human responsibility 475
 work rights and algorithmic management 380
 international environmental law 295n.5

- international human rights
 adjudication 266–67
 conventions and treaties 63, 69, 241, 266, 432
 courts 265, 268
 legal framework and instruments 93, 105–6,
 228–29, 324–25, 517–18, 520
 norms 37, 151
 obligations 312, 324
 standards 230, 243, 316, 324
see also international human rights law (IHRL)
- international human rights law (IHRL) 69, 85–
 86, 203–4, 356–57, 437–38, 519, 520, 570
 asylum rights 324–25, 328–30
 common risks and benefits 442, 451–52, 453–54
 consumer protection rights 410, 412, 423–24
 fair trial rights 268
 food rights 343–44
 healthy environment right 436
 religious freedom rights 69
 treaties 228–29
 women's rights and deepfake pornography 243
- international humanitarian law 327–29
- International Labour Organization
 (ILO) 254–55
- international law 198–202, 241, 328–30
 public 520
see also international human rights law (IHRL)
- International Lesbian, Gay, Bisexual, Trans and
 Intersex Association (ILGA) 233–34
- International Organisation for Migration
 (IOM) 323–24
- International Organisation for Standardisation
 (ISO) 214–15, 538n.45
- international treaties 104–6, 241
- internet shutdowns 88–89
- Internet of Things (IoT) 36, 349, 350–51, 361, 515
- interpretability 337
- IP address monitoring 34–35
- Iran 226
- Ireland: High Court 185
- iris-recognition 317, 324
- Italy 97, 185, 378, 535
 Modena University 407–8
- Iwasawa, Judge 199
- Japan 124–25, 130–31
- Jigsaw 224–25
- Johnson, A.M. 486–87
- joint controllership 124
- Jones, L. 198
- judges and judgments 488–90, 491–93,
 494, 497–98
see also data analytics in justice system and
 healthcare
- judicial mechanisms 520
- juries 497–98
- justice rights 4, 254, 409
- justice system *see* data analytics in justice system
 and healthcare
- justification clauses 560, 562, 565
- Kahneman, D. 268–69
- Kant, I. 467–68, 469
- Katrak, M. 9
- Kaye, D. 78–79
- Keller, Judge 268
- Kelly-Lyth, A. 376–77
- Kenya 154–55, 157–58, 298–99, 368
- key performance indicators 527
- Kleinberg, J. 271
- Klonick, K. 82n.28
- knowledge gaps 272, 285
- knowledge production 162–66
- Koen, L. 9
- Král, L. 7, 14–15
- Krupy, T. 550
- labelling 19
- Lancet Commission 401
- language 24, 89–90, 172–73, 231
 patterns 84–85
 understanding applications 391
see also natural language
- Latin America 508–9
- Latvia 315–16
- Laukyte, M. 10
- law enforcement and criminal justice 47–53
 automated facial recognition (AFR) 48–49
 predictive policing 50–52
 recidivism risk assessment 52–53
 social media monitoring 50
- lawfulness 125–26
- LCB v United Kingdom* 559
- LeCun, Y. 473–74
- legal personality 13–14, 65–66, 456–57, 458–70
 asking the right questions 458–61
 bodily integrity 464, 465
 Chinese room narrative 463–65
 consciousness 462–66, 470
 emotions 465–66
 empathy 465–66
 epistemology 462–66, 470
 functional superiority 466
 horizontal power relations between humans
 and AI 460
 metaphysics 462
 morals and ethics 461–62, 465–69, 470
 program 463–64, 466

- questions 463–64
 rationality and values 466, 467–68
 script 463–64
 story 463–64
 vertical power relations between humans and AI 460–61, 465–66, 469–70
- legal personhood 115
 legal positivism 490
 legal responsibility 496–97
 legal rights 481–82
 and natural (moral) rights distinction 478
 legal standards 167
 legality 64, 122
 legitimacy 64, 122, 268, 269–70
 lethal autonomous weapon systems (LAWS) 34, 484n.3
- LGBTQ+ rights 10, 89–90, 138, 147, 222–34
 censorship 226
 censorship mitigation 224–25
 conversion therapy 226–27
 death penalty 226–27
 deconstruction 231, 234
 depositions or court proceedings in hate crimes 224–25
 digital safe spaces 224–25
 filters 233–34
 forced ‘outings’ 226–27
 homophobia 224–25
 homosexuality criminalisation 226–27
 impact of AI 223–28
 imprisonment 226–27
 mental well-being in workplace 224
 negative aspects 223, 226–28, 234
 positive aspects 223, 224–25, 228
 queer theory 231, 234
 queerification 231–34
 queerifying 231, 232–34
 regulatory framework and discrimination protection 228–30
 repressive regimes 226–27
 safety 224–25
 torture 226–27
 see also sexual orientation and gender identity (SOGI) issues
- Li, S. 12
 liberty and security rights 7, 45–60, 113, 228–29, 446–47, 485
 emerging solutions 55–58
 international initiatives 56
 national legislation and case law 57–58
 regional proposals 56–57
 privacy, personal data protection and FRT 146
 security at risk 53–55
 see also law enforcement and criminal justice
- life cycle of AI: risks and remedies 7, 16–41, 527
 business understanding 17–18
 data preparation 18–24
 deployment 31–36
 evaluation 30
 modelling 25–30
 evaluation metrics and transfer learning 25
 unexplainable models 25–30
 requirements 37–40
 human rights risk assessment 37–39
 transparency 39–40
 life, right to 406, 475–76, 485, 557
 Lima, G. 475–76
 limitation techniques 8
 liquidity test 563–64
 Lisbon Treaty (2009) 405–6
 literary, scientific or artistic works including inventions 105
 Llewellyn, K. 268–69
 LMIC devices 402
Loomis v Wisconsin and COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software 52–53, 58, 113, 197, 271–72, 275–77, 278–79, 287–88
- McCulloch, W.S. 3
 McElroy, E. 369
 McEwan, I. 61
 McGill, J. 285
 machine learning (ML)
 asylum rights 319–20, 332
 common risks and benefits 445, 449–50
 data analytics in justice system and healthcare 281
 data preparation 18, 19, 24
 deployment 32–33
 disability rights 252
 effective remedy right and ADM 301–2
 fair trial rights 269–70, 271, 272–73, 275–76, 280
 gender-based discrimination 210
 health rights 387–88, 390–92, 395, 397–98, 400
 healthcare 4
 healthy environment right 426–27
 housing rights 368
 legal personality 464–65
 LGBTQ+ rights 223, 224–25, 233–34
 life cycle of AI: risks and remedies 16, 17
 modelling 25–26
 privacy and personal data protection 126, 129, 130, 131, 132–33, 138
 property rights and IPR 108–9, 114–15
 transparency 39–40
 work rights and algorithmic management 376

- McLeod Rogers, J. 550
 McNally, P. 473
 Maddocks, S. 238
 Majewski, L. 12
 Malaysia 88–89, 226
 Malgieri, G. 544–45
 malicious use of AI technology 7, 34–35, 454
 Malta: Digital Innovation Authority 554–55
 management-based approach 11, 56, 202–3, 206
 Mantelero, A. 548
 marginalised groups 9–10, 11, 14, 84, 85, 446–47, 570
 see also exclusion; vulnerable groups
 margin of appreciation 560, 563
Marsh v Alabama 87n.51
 Marx, J. 476, 479–80
 mass surveillance 136, 143, 146, 147, 169, 178, 447–48
 Mateescu, A. 371
 media and journalistic sources 79–80, 81
 media pluralism 77–78, 80
 Menéndez González, N. 9
 metaphysicals 480–81
 Microsoft 543–44
 AI for Earth programme 428
 micro-targeting of users for advertising purposes 88, 89
 migration management *see* asylum rights
 Mindstrong Health 389
 minorities *see* ethnic minorities, minorities and disadvantaged groups
 misclassification 252
 misconduct 116–18
 missing children and FRT use 97
 misuse and abuse of AI 36, 50, 60, 91–92, 545–46
 mitigation tools 13, 27, 134–35, 549
 mobile phone tracking 314–15
 model inversion 132
 Molbæk-Stensig, H. 10
 monitoring 7–8, 36, 37, 38–39, 375–76
 Moore, G. 531–32
 moral issues *see* ethical/moral issues
 Mufamidi, K. 9
 Myanmar 318
 Military Junta 157–58
 Rohingya genocide 84–85
 Nagel, T. 465
 National Emergency Management Agency 363–64
 national identity card system 157–58
 natural language 164, 224–25
 natural language processing (NLP) 25, 284, 391, 420–21
 natural (moral) rights 13–14, 477, 478–81
 nature of AI 542–43
 nature, rights of 477
 nearest neighbour models 25–26
 necessity principle 7–8, 64, 133–34, 145–46, 540, 546–48, 562
 negative obligations 15, 46–47, 85–86, 93, 121–22, 169
 negative rights 169, 451–52
 negative spirals of rights enjoyment 6
 Nepal 152–53
 Netherlands 50, 71–72, 97, 304–5, 335
 Amsterdam Algorithm Register 40
 CAS (Criminaliteits Anticipatie Systeem) 52
 childcare benefits and welfare benefits
 surveillance cases 223
 data protection authority (DPA) 145–46
 Impact Assessment for Human Rights in the Use of Algorithms 58
 ProKid 12-SI 22n.16
 smart cities and public law 512–13, 514
 SyRI (System Risk Indication) anti-fraud system 272, 502, 531–32
Toeslagenaffaire double nationality and tax fraud 454, 531–32
 Totta data lab for social assistance fraud prediction 300–1
 neural networks 25–26, 27, 108–9, 127, 226–27, 420–21, 430
 New Zealand 185, 521
 Court of Appeal 187–88
 YORST (Young Offending Risk Screening Tool) 52–53
 Nguyen, A. 371
 Nigeria 88–89
 Nightingale project 397
 NJCM (NGO) 272
 ‘noise’ 268–69
 non-discrimination 4, 9, 10, 11–12, 387
 asylum rights 323, 328, 329–30
 consumer protection rights 406
 data analytics in justice system and healthcare 293
 data preparation 19–21
 disability rights 248, 256, 259
 effective remedy right and ADM 294, 299–300, 303n.55, 304–5
 fair trial rights 266, 268–69, 271–72, 277–78
 freedom of expression 78–79
 health rights 393–94
 LGBTQ+ rights 226, 227–30
 nationality 334
 privacy and personal data protection 126

- privacy, personal data protection and FRT 141–42, 147, 149
 property rights and IPR 112
 religious freedom 62–63
 women's rights and deepfake pornography 240
 non-governmental organisations (NGOs) 321–22
 non-normativity 497–98, 499–500
 normative issues 453, 455–56, 485
 'Norms on the Responsibilities of Transnational Corporations and Other Business Enterprises with Regard to Human Rights' 518
 Norway Data Protection Authority 554–55
 Transparency Act 523
 nudging 165, 173
 object code 109–10
 object recognition 38
 objectification of physical features 143–44
 obligations immediate 357
 preventive 559
 procedural 133
 state 13, 370
 substantive 133
 to provide 74, 346–47
see also negative obligations; positive obligations
 Office of the United Nations High Commissioner for Human Rights (OHCHR) 69, 254–55, 525
 Omotubora, A. 11
 O'Neil, C. 271
 online dispute resolution (ODR) 12, 409, 565
 online learning 33
 online platforms freedom of expression 80, 84–85, 87, 88–89, 90
 private governance 81–82
 property rights and IPR 111–12
 workers 372, 383–84
 ontological approach 469, 479
 opacity *see* 'black box' and opacity
 openness 282–83
 operational records 39–40
 operator factors 96–97
 oppression 473, 474
 optimisation AI-technologies 11
 Optolexia 391–92
 Organisation of American States (OAS) 213
 Organisation for Economic Co-operation and Development (OECD) 203–4, 214, 436–37
 Data Protection Principles 124–25
 Guidelines for Fair Information Practice Principles (FIPPs) 123, 125
 Guidelines for Multinational Enterprises 523
 Guidelines on the Protection of Privacy and Transborder Flows of Personal Data 166
 National Contact Points 520
 organisational measures 96–97, 102
 organised crime 47–48
 Ortalda, A. 14–15
 Orwell, G. 68, 190
Osman v United Kingdom 559
 output layer 127
 OutRight Action International 233–34
 overadjustment/overcompensation 399
 overreliance 429
 oversight 56–57, 259, 324–25, 423
 AI decision-making limits 484–86, 488–89
 consumer protection rights 424
 freedom of assembly: biometrics and surveillance 102–3
 health rights 399–400
 work rights and algorithmic management 381–82
 Pakistan 152–53
 Papaléo Gagliardi, M. 9–10
 Paris Convention 105–6, 108
 Pasquale, F. 114–15
 password cracking 34
 patent filing 72
 patent protection 107–10
 peaceful enjoyment of one's possessions 107–8
 Penney, J. 98–99
 Percolata 373
 personal data protection asylum rights 323, 324–25, 337, 338–39
 common risks and benefits 445
 consumer protection rights 411–12, 418–19, 420–21
 data analytics in justice system and healthcare 281–83, 285, 290, 293
 data preparation 19, 23, 24
 deployment 32–33
 disability rights 257, 260–61
 effective remedy right and ADM 298–99, 301, 306
 food rights 352–53
 freedom of assembly: biometrics and surveillance 99–100, 102–3
 health rights 395, 396, 397, 398–99, 402
 human rights impact assessment (HRIA) 534–35

- personal data protection (*cont.*)
 property rights and IPR 112, 113
 public space, surveillance and right to be ignored 184, 189, 190
 regulatory sandbox 554–55, 566
 work rights and algorithmic management 371–72, 375–76, 381–82
see also privacy and personal data protection
- personalisation of content 81
- persons with disabilities (PWD) 364–65, 531–32
see also disability rights
- person(s) of interest 49, 94–97
- Peru 124–25, 129
- PetaBencana 363–64
- Phillips, Lord 187
- Pin, A. 9
- Pitruzzella, G. 77n.5
- Pitts, W. 3
- pluralism 7–8, 81, 422
- poachers 429
- Pokémon Go* 181–82
- Poland
 Random Allocation of Cases 40
 Supreme Administrative Court 40
 Tax Authority - Clearance Chamber ICT system 301
- police and law enforcement 7, 446–47
 abuse 98
 body-worn cameras 48
 practices, disproportionate 103
see also predictive policing
- policy team 527
- political beliefs 138
- political participation *see* privacy, political participation and FRT
- political rights *see* civil and political rights
- Pollicino, O. 77n.5
- Pop, A. 488–89
- Popova, A. 146
- pornography *see* women's rights and deepfake pornography
- positive obligations 9, 15, 452
 biometrics 93–94
 discrimination prevention 200–5, 206
 freedom of assembly: biometrics and surveillance 103
 freedom of expression 85–87, 90
 liberty and security rights 46–47
 privacy and personal data protection 121–22
 racial discrimination 206
- power asymmetry 173, 502, 505
- pre-emptive classification of user-generated content 83
- precautionary principle 15, 468, 469, 566
- predictive accuracy 420–21
- predictive analytics
 asylum rights 313–14, 315
 asylum rights and screening 331, 332, 333–34
 food rights 348–49, 350–51
 healthy environment right 429
 justice system and healthcare 283–84, 286, 291
 work rights and algorithmic management 373
- predictive maintenance 408–9
- predictive policing 416, 526–27
 human rights impact assessment (HRIA) 547–48
 law enforcement and criminal justice 50–52
 liberty and security rights 54–55, 56–57, 59
 location-based tools 51
 person-based tools 51
 racial discrimination 200
 smart cities and public law 502–3, 510, 511
- predictive systems
 fair trial rights 276–77
 versus public law 511–12, 514–16
see also predictive policing
- prescriptive analytics 283–84, 291
- prevention and mitigation system 569
- preventive obligations 559
- PricewaterhouseCoopers 224
- Prifti, K. 13
- Privacy Impact Assessment (PIA) 535n.26
- privacy and personal data protection 12, 121–35
 accountability 133–34
 accuracy 130–31
 applicability 123–25
 data minimisation and storage limitation 129–30
 lawfulness and fairness 125–26
 production of and control over data 162–76
 ambience and opacity 173
 ideal-typical models of data regulation 166–70
 ideal-typical models of data regulation, failure of 170–74
 knowledge production 162–66
 nudging 165, 173
 power imbalance 173
 practical limits 173
 purpose limitation 128–29
 security, integrity and confidentiality 131–33
 transparency 127–28
see also privacy, personal data protection and FRT
- privacy, personal data protection and FRT 136–49
- blanket and indiscriminate retention 143–44, 147–48

- child, rights of 147–49
 equality and non-discrimination 147
 freedom of assembly and association 146
 FRT 137–38
 human dignity 143–44
 privacy rights 136–37
 security 144–46
 privacy, political participation and FRT 150–61
 informational privacy, moving beyond 159–61
 prevailing conception of privacy 152–56
 Africa: data protection 154–55, 156–58, 159
 Asia: informational conception 152–53, 157–61
 South America: informational
 privacy 155–59, 161
 surveillance state and digital
 authoritarianism 156–58
 privacy and privacy rights 4–5, 8–9, 11–12, 167
 abuse 10, 132
 asylum rights 323, 324–25, 333–34, 337
 behavioural privacy 375–76
 business and human rights (BHR) and private
 sector responsibility 517
 common risks and benefits of AI 444–48,
 449–50, 452
 consumer protection rights 407–8, 411–
 12, 420–21
 data analytics in justice system and
 healthcare 281–83, 285, 290, 293
 data preparation 19, 23–24
 decisional privacy 160n.93
 deployment 32–34, 36
 disability rights 252–53, 257
 effective remedy right and ADM 298–99
 food rights 352–53, 354
 freedom of assembly: biometrics and
 surveillance 99–100, 102
 freedom of expression 78–79, 89
 health rights 402
 human rights impact assessment
 (HRIA) 534–35
 informational privacy 160n.93, 375–76
 LGBTQ+ rights 224–27, 228–29, 234
 local privacy 160n.93
 property rights and IPR 112, 113
 public space, surveillance and right to be
 ignored 177, 181–84
 regulatory sandbox 554–55, 557, 559, 564–65
 religious freedom right 68, 71–72
 women's rights and deepfake
 pornography 236, 237, 239, 240, 241,
 243, 246, 247
 work rights and algorithmic management 371–
 72, 375–76, 377–78, 380
- see also* privacy and personal data protection;
 privacy, personal data protection and
 FRT; privacy, political participation
 and FRT
 private media and information platforms 7–8
 private sector 298–301, 302, 304–5
 private sector responsibility *see* business and
 human rights (BHR) and private sector
 responsibility
 procedural equality principle 296–97
 procedural fairness/fair treatment 4–5, 320
 procedural mechanisms 46, 504–5
 procedural obligations 133
 procedural restraints 39
 procedural rights 281, 295–97, 503–4
 procedural safeguards 57–58, 297–98
 product team 527
 product unsafety 12
 profiling 89, 113–14, 125–26, 250–51
 asylum rights 322–23
 data analytics in justice system and
 healthcare 287–88, 290
 disability rights 252–53, 256
 group 165
 privacy and personal data protection 166–67
 racial 156–57
 property rights and intellectual property
 rights 8, 104–18, 449–50
 consequences 110–15
 copyright protection 106–8, 109–10
 exceptions and limitations 115–17, 118
 implementation of rights and relevance
 to AI 104–10
 intellectual property as human right 106–8
 international treaties 104–6
 originality 109–10
 patent protection 107–10
 policy proposals for improvements 117–18
 regulatory sandbox 557
 robot rights/human responsibility 477
 tangible property 107–8
 trade secrets protection 108, 109–10, 113–17
 proportionality 7–8, 570
 freedom of assembly: biometrics and
 surveillance 92, 99, 100
 human rights impact assessment
 (HRIA) 540, 546–48
 privacy and personal data protection 122,
 126, 134
 privacy, personal data protection and
 FRT 144–46, 148
 property rights and IPR 114–15
 regulatory sandbox 560, 562–63
 smart cities and public law 504–5

- proprietary rights 104, 106–7, 113
 protect, duty to 7, 69–75, 346–47, 454, 519
 protected value
 may be taken into account 20–21
 shall be taken into account on a equal basis 21
 shall not be taken into account 21
 provide, obligation to 346–47
 proxies 144, 209, 256, 376, 417, 418–19
 discrimination 210
 pseudonymisation 23–24, 132, 395–96
 public control of governmental AI 40
 public interest
 common risks and benefits 453–56
 consumer protection rights 406–7
 freedom of assembly: biometrics and surveillance 92
 health rights 396–97
 privacy, personal data protection and FRT 142–43, 146
 property rights and IPR 105–7, 115, 116–18
 public law 13, 14
 international 520
 see also smart cities and public law
 public order 50
 public safety 68, 560
 public sector 152–53, 223, 325, 383–84, 423–24
 effective remedy right 298–99, 300, 301, 302, 304
 smart cities 503–4, 512–13, 514
 public space, surveillance and right to be ignored 177–91
 anonymity 186–87, 188–91
 benefits of AI in public spaces 179–81
 confidentiality 187–89, 190–91
 going beyond individualism 189–90
 going beyond privacy 184–89
 privacy issues in public spaces 181–84
 public watchdogs 80
 purpose of AI 542–43
 purpose limitation principle 134–35, 306, 396n.101
- Quemby, A. 10
 Quinn, P. 544–45
 Quintavalla, A. 13
- R (Bridges) v South Wales Police and AFR*
 Locate 58, 180, 185, 186, 200–2, 203, 204–5, 206
 Rachlinski, J.J. 286–87
 racial discrimination 9–10, 195–206, 287
 AI decision-making limits 489–90
 business due diligence and positive duties for discrimination prevention 202–5, 206
- business and human rights (BHR) and private sector responsibility 517–18
 direct 195, 200
 indirect 195, 199–200
 international law and prevention obligations 198–202
 legal personality 460
 positive obligations 200–2, 206
 prevalence in AI 195–98
 racial equality 10
 racial profiling 156–57
 Radio-Frequency Identification (RFID) tags 533
 Ranchordás, S. 14
 Ranking Digital Rights 528–29
 Raso, F. 223
 Raz, J. 491–92n.36
 re-politicisation 7–8
 real-life experimentation with AI *see* regulatory sandbox
Reason 61–62
 reasoned judgment rights 297–98, 302–3, 304–7
 rebound effect 429
 recidivism and recidivism risk assessment 52–54, 56–57, 58, 59, 416
 recommender systems 80–81, 83, 85, 88, 89–90, 421–22
 recruitment 207, 210, 213, 219, 299–300, 376–77, 443–44, 460
 rectification 130–31
 Reddit 239, 244–45
 redress strategies 7, 13
 refugee screening *see* asylum rights and screening
 Regan, T. 477
 regulatory sandbox 15, 456–57, 551–66
 concept and impact on development 553–55
 fundamental rights concerns 556–58
 proactive measures for remedy and relief 563–65
 violation prevention 559–63
 relational rights 6, 476
 religious artefacts 66–67, 72–73, 74
 religious freedom 7, 61–75
 freedom to have a religion or belief (inner freedom or *forum internum*) 63–65, 66, 68, 75
 freedom to manifest one's religion or belief externally (outward freedom or *forum externum*) 64, 66
 fulfil, duty to 74
 protect, duty to 69–75
 respect, duty to 63–68, 71, 75
 remote workforce 372
 RentCheck 360–61
 respect, duty to 7, 63–68, 71, 75, 346–47

- restrictive measures 88–89
 resumé-mining tools 257–58
 retrospective review 39
 reuse of pre-trained models *see* transfer learning
 revisability 39
 Revised European Social Charter (Social Charter) 356, 358, 359
 Reyes, R. 11
 right answer thesis 274–75, 280
 right to pursue a chosen economic activity 110–11
 rights balancing 144
 Rio Declaration 295n.5
 risk
 acceptance 548
 assessment 11–12, 102, 202, 224, 538–39
 see also human rights risk assessment
 association 450–51
 autonomy 450–51
 financial 563–64
 indicators 502
 network 450–51
 -scoring systems 502–3
 structural 13, 443–46, 453, 455, 456
 see also structural risks
 robot rights/human responsibility 471–83
 analysis of rights 476, 478
 claims 476
 consciousness 480–81
 correlated duties 476
 currency of rights 476
 differentiation of rights 478–82
 epistemology 480–81
 getting rights right 474–76
 getting rights wrong 472–74
 immunities 476
 incidents 476
 legal rights 481–82
 metaphysical properties 480–81
 moral status 479
 natural (moral) rights, and legal rights
 distinction 478
 natural (moral) rights 478–81
 ontological condition 479
 powers 476
 privileges 476
 relational rights 476
 robots 13–14, 61–62, 248, 251–52, 348–49, 445, 473–74
 see also robot rights/human responsibility
 robustness 381–82
 Rome Declaration on World Food Security 343–44
 Royal Free London NHS Foundation 397
 Ruggie, J. 518–19
Ruiz Rivera v Switzerland 268
 rule of law test 64
 Russia 48–49, 146
 freedom of assembly: biometrics and surveillance 91
 LGBTQ+ rights 224–25, 226
 Orwell FRT in schools 179–80
 regulatory sandbox 554–55, 564–65
 women's rights and deepfake pornography 238
 Rwanda: Tutsi Genocide 318
 Ryberg, J. 271
S and Marper v the United Kingdom 143–44, 147–48
 safeguards 101–3
 Salyzyn, A. 285
 sanctions 60, 93, 98–99, 213
 Sanders, J.W. 494–95
 sanitation 11, 367–69
 Santamaría Echeverría, E. 12
 Saudi Arabia 226
 ‘Sophia’ humanoid robot and honorary citizenship 473–74
 scanner software 142, 315
 Scassa, T. 102
 Schauer, F. 98–99
 Schermer, B. 533
 Schütte, B. 12
Schwabe 100
 Schwelb, E. 198
 scope and objectives of AI 4–6, 542–43
 scoring algorithms 113–14
 screen reading devices 251
 screening *see* asylum rights and screening
 search engines 81–82
 Searle, J.R. 464
 second generation rights 6, 107–8, 453
 secrecy 116–17, 134–35, 271–72
 see also confidentiality; trade secrets protection
 security 5, 10, 406
 consumer protection rights 407–8
 freedom of assembly: biometrics and surveillance 102
 privacy and personal data protection 131–33
 privacy, personal data protection and FRT 142–43, 144–46, 149
 smart cities and public law 510
 threats 47–48, 50, 314
 see also liberty and security rights
Segerstedt-Wiberg 99–100
 Selbst, A.D. 223

- self-autonomy 419–20
 self-binding principles 214–15
 self-determination, informational 168
 self-regulatory measures 70, 86, 259, 382–83
 sensitive data 33
 disability rights 259
 emotions and demographics 95–96
 freedom of assembly: biometrics and surveillance 99–100
 freedom of expression 79
 health rights 395–96, 399
 LGBTQ+ rights 226–27
 privacy, personal data protection and FRT 139–40
 public space, surveillance and right to be ignored 177
 Sensity AI (formerly Deeptrace Labs) 236
 sentiment analysis 256, 379
 Seo-Young Chu 471
 separation theorem 467
 sexual orientation and gender identity (SOGI)
 issues 10, 224–25, 226–27, 228–29, 230
 shadow-banning 85, 89–90
 Shahid, M. 10
 Sharkey, N. 471
 Sibony, O. 268–69
 Singapore 88–89
 Advisory Council on Ethical Use of AI and Data 337
 Singer, P. 477
 SkinVision 387–88
 Slove, D.J. 152
 smart cities and public law 14, 182–83, 447–48,
 456–57, 501–16
 administrative law 511–12, 513, 514–16
 civil law 511–12, 515–16
 definition of smart city 505–6
 do-it-yourself public services 512–14
 exclusion 508–11, 515
 local law 515
 narrative, smart city as 507–8
 power asymmetry 502, 505
 predictive policing 502–3, 510, 511
 predictive systems versus public law 511–12, 514–16
 process, smart city as 508
 product, smart city as 507, 508
 public services 501–2
 security 510
 smart citizenship 509
 smart grids 502–3
 social welfare fraud enforcement 502–4, 512, 514–15
 strategy, smart city as 506–7, 508
 systems versus individual human flaws 514–15
 vulnerable citizens 502, 503–4, 510, 512–15
 smart farming 11, 344, 348–53
 smart grids 502–3
 smart home devices 251–52, 253
 smart homes 361, 369–70
 smart riot surveillance systems 91
 smart use cases 13
Smith Kline v the Netherlands 107–8
 SMM software 54
 Šmuclerová, M. 7, 14–15
 social conformity effect 98–99
 social contract theory 468–69
 social crediting 34–35
 social media and social networks
 business and human rights (BHR) and private sector responsibility 528
 freedom of assembly: biometrics and surveillance 97
 freedom of expression 80–82, 84–86, 87, 88–89, 90
 law enforcement and criminal justice 50
 LGBTQ+ rights 224–25
 privacy and personal data protection 128
 property rights and IPR 111–12
 social rights *see* economic, social and cultural rights
 social security rights 300–1
 social sorting 157–58
 social welfare fraud enforcement 502–4, 512, 514–15
 soft law
 business and human rights (BHR) and private sector responsibility 519
 disability rights 258
 fair trial rights 277–78
 gender-based discrimination 211–12, 213–15, 218–19
 work rights and algorithmic management 382–83
 Software Directive 109–10
 Solove, D. 121
 Somalia 314, 327–28, 332
 Somayajula, D. 11
 source code 109–10
 South Africa 154–55, 390
 Cmore system 52
 Promotion of Equality and Prevention of Unfair Discrimination Act (PEPUDA) 204–5, 206
 South America 155–59, 161
 South Korea 152–53, 508
 Personal Information Protection Act 535

- Spain 383–84
 Labour Code 291
- Spano, R. 268
- Sparrow, R. 486–87
- speech recognition 33
- speech-to-text algorithms 251
- Sri Lanka 153n.29, 153n.31
- Stamhuis, E. 15
- state obligations 13
- statistical modelling 489–90
- Stefano, V. de 383
- Stockholm Declaration 431
- stocktaking 6
- storage limitation 129–30, 134–35
- stress-testing 527–28
- structural inequalities 369–70
- structural oppression 254
- structural risks 13, 443–46, 453, 455, 456
- structuralism approach 202–3
- Sturm, S. 202–3
- Sub-Commission on Prevention of Discrimination and Protection of Minorities 198–99
- substantive aspects 46, 456
- substantive human rights 5, 6, 294, 295–96, 298, 299–300, 306–7
- substantive obligations 133
- suitability requirements 145–46
- Sunstein, C.R. 81, 268–69
- supervised learning 543–44
- surrogate models 27
- surveillance 7, 8–9
 asylum rights 314–15
 control 11
 freedom of expression 79
- Netherlands welfare benefits surveillance cases 223
- physical 447–48
- privacy, personal data protection and FRT 143–44
- privacy, political participation and FRT 150–51, 152
- property rights and IPR 113
- racial discrimination 200–1
- religious freedom rights 68, 71–72
- smart cities and public law 510
- work rights and algorithmic management 375–76, 385
- see also* freedom of assembly; biometrics and surveillance; mass surveillance; public space, surveillance and right to be ignored
- surveillance capitalism 78–79, 89
- surveillance state 156–58
- Sweden 48–49, 67, 97
 data protection authority (DPA) 141–42
 Ombudsperson 66–67
- Switzerland 267
 GeoMatch 335
- synthetic data 23, 24
- Sypniewski, A. 268–69
- Syria 327–28
- systematic description of AI system 542–46
- technical affordances 96–97
- technical restraints 39
- technical specifications 17–18
- Telegram app 240
- Temperman, J. 7, 13
- Tencent's WeChat 226
- terms and conditions 82–83
- terrorism 47–48, 101, 159
see also cyberattacks
- Tesla 24
- test bed 27
- testing data 18, 19, 20, 25, 30, 132, 278–79
- text classification or language translation 25
- Thailand 153n.31, 390
- thermal scanners 142
- third countries 72
- third generation rights 6, 12, 453
- third parties 45, 86–87, 133, 142, 528
- Thompson Reuters' *Litigation Analytics* 285
- Tiefensee, C. 476, 479–80
- Tokyo Nutrition for Growth (N4G)
 Summit 343–44
- Tomada, L. 8
- Toronto Declaration 423–24
- tort liability 470
- torture, degrading, and inhuman treatment 265, 544
- Total 522–23
- totalitarian regimes 150–51, 159, 169
see also digital authoritarianism
- trade secrets 108, 109–10, 113–17, 301–2, 420–21
- training data 9–10, 19, 20, 25, 95, 398–99, 416–17
 asylum rights and screening 335–36, 338
 disability rights 248, 253–54, 259
 gender-based discrimination 208–9, 210, 217–18
- privacy rights 126, 128, 132
- property rights and IPR 108–9, 114–15
- work rights 373, 376
- transfer learning 25, 39–40
- transfer phase 7
- transnational value chains 517–18

- transparency 7, 39–40
 asylum rights 316–17, 320, 322, 323, 324–25
 asylum rights and screening 336, 338–39
 business and human rights (BHR) and private sector responsibility 521, 522–23, 526–27, 528–29
 common risks and benefits 444, 445–46, 454–55, 456
 consumer protection rights 419–22, 424
 data analytics in justice system and healthcare 281, 282–83, 285, 288–89
 development 28, 39
 effective or qualified 114–15, 117–18
 fair trial rights 268, 271, 272, 277–79
 food rights 347–48, 349–50, 354
 freedom of assembly: biometrics and surveillance 102–3
 freedom of expression 78, 88
 gender-based discrimination 213, 219
 health rights 395, 402
 healthy environment right 429, 434
 human rights impact assessment (HRIA) 537–38
 human rights risk assessment 37, 38–39
 liberty and security rights 55, 56–57, 59
 modelling 29, 30
 operational 28, 29–30, 39
 privacy and personal data protection 125, 127–28, 134–35
 property rights and IPR 104, 111–12, 113, 114–17
 psychological factor 486–87
 religious freedom, right to 74
 smart cities and public law 512
 work rights and algorithmic management 381–82
- Treaty on the Functioning of the European Union (TFEU) 122–23
- Trevor Project 224, 233–34
- trust 40, 282–83, 285, 292
- trustworthiness 563, 566
- truth detection 330
- Turing, A./Turing Test 3, 484n.4
- Turner, J. 481–82
- Tutt, A. 114–15
- Twitter 244–45, 363–64
 API 363
 Tay chatbot and discriminatory and offensive messages 543–44
- Tyrer v the United Kingdom* 532
- Tyson Foods Inc.* 351
- Uber 374–75, 379–80, 415–16
- Uganda 156–58, 390, 522–23
- Ukraine 327–28, 343–44
- unbalanced data 25, 33
- unexplainability 12, 25–30
 black box systems restrictions 27–30
 health rights 394–95, 402
 healthy environment right 429
 mitigation methods 27
- unforeseeability and unpredictability 129
- unintelligibility 127–28
- United Arab Emirates (UAE) 226
- United Kingdom 125, 130
 AI decision-making limits 496n.58
Airey v Ireland 559
 Artificial Intelligence in Healthcare sandbox 554–55
 asylum rights 320–21, 331
 automated facial recognition (AFR) 48–49
 British Sandbox 554
 business and human rights (BHR) and private sector responsibility 522
- Business, Human Rights and Environment Act 522–23
- Centre for Data Ethics and Innovation 337
- Court of Appeal 187–88
- Data Commissioner's Office (ICO) 139
- Data Protection Act 1998 397
- data protection authority (DPA) 139
- Digitalisation at Work 383–84
- Equality Act 2010 201–2, 334
- Financial Conduct Authority (FCA) 554
- HART (Harm Assessment Risk Tool) 52
- Home Office 334
 'hostile environment' policy 335–36
- Human Fertilisation and Embryology Authority (HFEA) 400
- human rights impact assessment (HRIA) 535
- Independent Chief Inspector of Borders and Immigration 334
- Independent Workers of Great Britain 384
- Information Commission 397
- London Metropolitan Police 95
- Modern Slavery Act 2015 521, 522–23, 526–27
- National AI Strategy 58
- Online Harms Bill 523
- Online Safety Bill 523n.22
- predictive policing 51
- public space, surveillance and right to be ignored 183, 185
- R (Bridges) v South Wales Police and AFR Locate* 58, 180, 185, 186, 200–2, 203, 204–5, 206
- racial discrimination 204–5
- work rights and algorithmic management 382

- United Nations 255, 324–25, 343–44
 gender-based discrimination 207–8, 210,
 211–13, 215–17, 218–19, 220–21
- United Nations Charter 3, 207–8
- United Nations Children's Fund (UNICEF) 254–55
- United Nations Commission for Refugees (UNHCR), 'Project Jetson' 314, 332
- United Nations Committee on Economic, Social and Cultural Rights (CESCR) 254–55, 364–65, 368
- United Nations Convention on the Law of the Sea 295n.5
- United Nations Convention Relating to the Status of Refugees (1951) 318, 328–29
- United Nations Convention on the Rights of Persons with Disabilities (CRPD) 10, 249–50, 254–58, 260
- United Nations Educational, Scientific and Cultural Organization (UNESCO) 254–55
- Institute for Information Technologies in Education 79–80
- Recommendation on the Ethics of Artificial Intelligence 56
- United Nations General Assembly (UNGA) 3, 198–99, 216–17, 330, 405–6
 Declaration on Territorial Asylum (1967) 329
- United Nations Global Compact 436–37, 519
- United Nations Guidelines for Consumer Protection (UNGCP) 405–6
- United Nations Guiding Principles on Business and Human Rights (UNGPs) 86–87, 213–14, 230, 244, 305, 423–24, 436–37, 454, 534–35
 business and human rights (BHR) and private sector responsibility 518–21, 522, 524–27, 528, 529–30
- United Nations High Commissioner for Human Rights (UNHCHR) 99, 101, 156–57, 207–8, 216–17, 258, 261, 432–33
- United Nations High Commissioner for Refugees (UNHCR)
 asylum rights 317, 318, 321, 323–24, 335
 Brazil 332
 Refugee Status Determination (RSD) 319, 320
- United Nations Human Rights Committee 216–17, 269
- United Nations Human Rights Council (HRC) 26, 47, 63–64, 102–3, 213–14, 216–17, 423–24, 528–29, 531–32
- Advisory Committee 56
- General Comment on the Freedom of Peaceful Assembly 98
- healthy environment right 425, 431, 432–33, 436–37
- 'Protect, Respect, Remedy Framework' (2008) 518–19
- Resolution 48/13 434
- United Nations News website 196
- United Nations Office for the Coordination of Humanitarian Affairs (OCHA) 313–14
- United Nations Principles of Responsible Investments 519
- United Nations Protocol Relating to the Status of Refugees (1967) 328–29
- United Nations Refugee Agency 311
- United Nations Special Rapporteur 213–14, 432–34
- United Nations Special Rapporteur on Adequate Housing 355–56, 359–60, 362
- United Nations Special Rapporteur on Extreme Poverty and Human Rights 355–56
- United Nations Special Rapporteur on the Right to Food 347
- United Nations Special Rapporteur on Violence against Women 244
- United Nations Sub-Commission on Promotion and Protection of Human Rights 518
- United Nations Sustainable Development Goals (SDGs) 207–8, 343–44, 347
- United States
 AI decision-making limits 484n.2, 489–90
 asylum rights 315, 331
 automated facial recognition (AFR) 48–49
 Biometric Information Privacy Act (BIPA) 139, 143
 Border Patrol (USBP) agents 315
 California Transparency in Supply Chains Act 2010 521–22
 Capitol Hill rioters (2021) 69, 97
 Center for Disease Control and Prevention (CDC) 249–50
 Chicago police and Strategic Subject List software 51
 Chicago University 397–98
 Clearview AI 49, 97, 179–80
Connecticut Fair Housing Center v CoreLogic Rental Property Solutions 360
 Constitution - First Amendment 77, 87
 Constitution - Fourth Amendment 122
 consumer protection rights 415–16
 critical care resource allocation 197
 data analytics in justice system and healthcare 285, 286n.19
 Dodd-Frank Act 2010 521–22
 Fair Housing Act 360
 fair trial rights 268–69, 280

- United States (*cont.*)
- Food and Drug Administration (FDA) 387–88, 395
 - food rights 350–51
 - freedom of expression 88–89
 - gender-based discrimination 213
 - Health Insurance Portability and Accountability Act (HIPPA) 397–98
 - health rights 397
 - housing rights 364–65
 - Illinois Biometric Information Privacy Act (BIPA) 137
 - Immigration and Customs Enforcement Agency (ICE) 315–16
 - LGBTQ+ rights 232–33
 - Loomis v Wisconsin* and COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software 52–53, 58, 113, 197, 271–72, 275–77, 278–79, 287–88
 - Los Angeles Police Department (LAPD) Data-Informed Community-Focused Policing (DICFP) 51
 - Los Angeles Police Department (LAPD)
 - Operation Laser and Palantir software 51 - Massachusetts and Washington DC and court orders for AFR 57–58
 - National Security Agency (NSA) ‘PRISM’ project 447–48
 - New York City 36, 208, 213
 - New York City - Public Oversight of Surveillance Technology (POST) Act 57–58
 - Northern Virginia data centres 429–30
 - Oregon Department of Justice and Digital Stakeout SMM tool 50
 - predictive policing 51
 - privacy and personal data protection 130–31
 - privacy, political participation and FRT 157–58
 - Product Liability Law 415–16
 - public space, surveillance and right to be ignored 181, 184, 185
 - racial discrimination 5, 196
 - San Francisco AFR ban 57–58
 - San Francisco SOGI and LGBTQ+ 224–25
 - Santa Cruz predictive policing ban 57–58
 - state action doctrine 87
 - Supreme Court 184–85, 283–84
 - Wholefoods 377–78
 - Universal Declaration of Human Rights (UDHR) 3, 11
 - asylum rights and screening 327–29
 - effective remedy right 295
 - food rights 344
 - freedom of assembly 93
 - freedom of expression 76
 - gender-based discrimination 220
 - housing rights 356–57
 - human rights impact assessment (HRIA) 534–35
 - legal personality 459.n.6
 - privacy rights 121–22, 167
 - property rights and IPR 105
 - religious freedom 64
 - robot rights/human responsibility 475
 - unpredictability 133–34
 - unsupervised learning 543–44
 - Uruguay 125–26
 - user control 277–78
 - user input model 231–32
 - user requirements 17–18, 22, 37, 41
 - user stories 17
 - validation 30
 - van der Sloot, B. 9
 - Venezuela 157–58
 - verification 30
 - verification/authentication (one-to-one) function 138, 145–46
 - Videmo 360 face recognition system 95–96
 - video lie detector 315–16
 - video screening 257–58
 - video-sharing platform services 70–71
 - violence 69, 517–18
 - gender-related 243
 - incitement 528
 - mob 69
 - sexual 236 - virtual reality (VR) games 253
 - visual data 379
 - vital interest of data subject justification 125
 - voice recognition 21, 248, 251, 254, 330, 331, 332–33
 - voluntary accessibility commitments 258–59
 - voluntary measures 259
 - voting rights 475–76
 - vulnerable groups 132, 446–47
 - biases 56
 - business and human rights (BHR) and private sector responsibility 527–29
 - freedom of assembly: biometrics and surveillance 98
 - liberty and security rights 47
 - religious freedom rights 70
 - smart cities and public law 502, 503–4, 510, 512–15

- Wachter, X. 113–14
 Warren, S.D. 177
 Warthon, M. 7–8
 watchlists 48
 Watrix 298
 wearable devices
 robotic exoskeletons 251–52
 smart wristwatches 375–76
 Wenar, L. 476, 478–79
 white box models 27–28
 will theory 477–78
 Williams Institute 233–34
 Williams, J.C. 197
 Winterson, J. 61
 Woebot 389
 women's rights and deepfake
 pornography 235–47
 anonymous sources 242
 anonymous women 236
 celebrity deepfakes 236, 244–45
 consent, obligatory 246
 deepfake technology and harms 239–41
 discrimination, gender-based 236–39,
 241, 247
 dubbing 235
 economic damage 243
 entertainment 241–42
 financial gain 241–42
 gender oppression 237
 intent 242
 intimacy, violation of 238
 legal framework 236
 mitigation strategies 236
 moral and psychological damage 243
 negative consequence 247
 non-consensual dissemination of
 images 237–38, 241, 245, 247
 objectification and mediation of female
 bodies 237, 238
 political deepfakes 238
 privacy violation 237, 239, 240, 241, 243,
 246, 247
 rape culture 238
 realism of images 239
 regulatory framework and
 recommendations 241–46
 reputational damage 240
 revictimisation 242
 serious harm 242
 sexual abuse, image-based 237–38, 245–46
 sexual (gendered) violence 236, 237, 238–39,
 241, 243, 246
 sexual gratification 241–42
 Wood, A. 373
 Word2Vec 210
 work rights 11–12, 257–58, 281–82, 353, 416,
 444–48, 517
 work rights and algorithmic management 371–85
 anticipatory compliance practices 380
 collective bargaining 372, 383–85
 digital revolution at work 372–75
 direction of workers 373
 disability 376–77, 382
 discipline and termination 374–75
 employability score 373
 ex ante approach 381–84
 fair and just working conditions, right to 380
 ‘gig economy’ 372
 human rights dimensions 375–80
 occupational health and safety 380
 platform workers 372, 383–84
 qualitative dimension 371
 recruitment 376–77
 remote workforce 372
 unionisation risk 377–78
 Working Group on Transnational
 Corporations 518
 World Food Programme (WFP) 317
 World Food Summit (WFS) (1996) 343–44
 World Health Organisation (WHO) 248n.1,
 249–50, 367, 386n.1
 Wu, T. 83n.31
X and the Church of Scientology v Sweden 66–67,
 73–74
X and Y v the Netherlands 559
 Xenidis, R. 8
 Yam, J. 544–45
 Yeung, K. 101, 548
 Yogyakarta Principles 230
 YouTube 211, 389
 Zarra, A. 10
 Zencity 50
 Zillow - Zestimate program 298
 Zimbabwe 157–58
 Zornetta, A. 8–9
 Zuboff, S. 179
 Zwart, T. 269–70

