

Doctoral thesis

Governing Machine Learning that Matters

Michael Veale

2019

A dissertation submitted in partial fulfilment of the
requirements for the degree of

*Doctor of Philosophy in Science, Technology, Engineering and
Public Policy*

Department of Science, Technology, Engineering
and Public Policy (STeAPP)
University College London

84,570 words

2019

Declaration of Authorship

I, Michael Veale, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Declaration of Integrated Publications

This section acknowledges the integration of work of the Author into the different sections of this thesis. All work integrated into the thesis was undertaken during the period in which the Author was registered with the University as working towards the doctorate. The work is acknowledged here as, since publication, other researchers have responded to and engaged with these works in this fast-moving field, and this thesis represents both a statement of the original arguments and findings in those works, as well as a partial response to the research field as it stands at the time of submission.

Chapter 1, Hello, World!, includes some content from the following articles:

1. Vasilios Mavroudis and Michael Veale, 'Eavesdropping Whilst You're Shopping: Balancing Personalisation and Privacy in Connected Retail Spaces' in *Proceedings of the 2018 PETRAS/IoTUK/IET Living in the IoT Conference* (IET 2018) DOI: 10/gffng2;
2. 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 L. & Tech. Rev.* 18 DOI: 10/gdxthj;
3. Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao and Nigel Shadbolt, 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'18)* (ACM 2018) DOI: 10/cvcp

Chapter 2, The Law of Machine Learning?, draws upon and extends the following articles:

1. Michael Veale, Reuben Binns and Lilian Edwards, 'Algorithms That Remember: Model Inversion Attacks and Data Protection Law' (2018) 376 *Phil. Trans. R. Soc. A* 20180083 DOI: 10/gfc63m;

-
2. Michael Veale, Reuben Binns and Jef Ausloos, 'When data protection by design and data subject rights clash' (2018) 8(2) *International Data Privacy Law* 105 DOI: 10/gdxthh;
 3. Michael Veale and Lilian Edwards, 'Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling' (2017) 34(2) *Comput. Law & Secur. Rev.* 398 DOI: 10/gdhrtm;
 4. 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 L. & Tech. Rev.* 18 DOI: 10/gdxthj;
 5. Lilian Edwards and Michael Veale, 'Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?' (2018) 16(3) *IEEE Security & Privacy* 46 DOI: 10/gdz29v.

Chapter 3, *Data Protection's Lines, Blurred by Machine Learning*, draws upon and extends the following articles:

1. Michael Veale, Reuben Binns and Jef Ausloos, 'When data protection by design and data subject rights clash' (2018) 8(2) *International Data Privacy Law* 105 DOI: 10/gdxthh
2. Michael Veale, Reuben Binns and Lilian Edwards, 'Algorithms That Remember: Model Inversion Attacks and Data Protection Law' (2018) 376 *Phil. Trans. R. Soc. A* 20180083 DOI: 10/gfc63m
3. Michael Veale and Lilian Edwards, 'Better seen but not (over)heard? Automatic lipreading systems and privacy in public spaces' [2018] Presented at PLSC EU 2018

Chapter 4, *Coping with Value(s) in Public Sector Machine Learning*, draws upon and extends the following articles:

1. Michael Veale, 'Logics and Practices of Transparency and Opacity in Real-World Applications of Public Sector Machine Learning' in *Presented at the 4th Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT/ML 2017), Halifax, Nova Scotia, Canada, 2017* (2017) (<https://arxiv.org/abs/1706.09249>)
2. Michael Veale, Max Van Kleek and Reuben Binns, 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'18)* (ACM 2018) DOI: 10/ct4s

-
3. Michael Veale and Irina Brass, 'Administration by Algorithm? Public Management meets Public Sector Machine Learning' in Karen Yeung and Martin Lodge (eds), *Algorithmic Regulation* (Oxford University Press 2019)

Chapter 5, Unpacking a tension: 'Debiasing', privately, draws upon and extends the following articles:

1. Michael Veale and Reuben Binns, 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data' (2017) 4(2) *Big Data & Society* DOI: 10/gdcfnz
2. Michael Veale and Irina Brass, 'Administration by Algorithm? Public Management meets Public Sector Machine Learning' in Karen Yeung and Martin Lodge (eds), *Algorithmic Regulation* (Oxford University Press 2019)

Other publications produced during the course of this thesis and related to its subject matter but not integrated into the document include:

1. Max Van Kleek, William Seymour, Michael Veale, Reuben Binns and Nigel Shadbolt, 'The Need for Sensemaking in Networked Privacy and Algorithmic Responsibility' in *Sensemaking in a Senseless World: Workshop at ACM CHI'18, 22 April 2018, Montréal, Canada* (2018) (<http://discovery.ucl.ac.uk/id/eprint/10046886>)
2. Michael Veale, Reuben Binns and Max Van Kleek, 'Some HCI Priorities for GDPR-Compliant Machine Learning' in *The General Data Protection Regulation: An Opportunity for the CHI Community? (CHI-GDPR 2018). Workshop at ACM CHI'18, 22 April 2018, Montréal, Canada* (2018) (<https://arxiv.org/abs/1803.06174>)
3. Michael Veale, *Data management and use: case studies of technologies and governance* (The Royal Society and the British Academy 2017)
4. Niki Kilbertus, Adria Gascon, Matt Kusner, Michael Veale, Krishna P Gummadi and Adrian Weller, 'Blind Justice: Fairness with Encrypted Sensitive Attributes' in *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)* (2018) (<http://proceedings.mlr.press/v80/kilbertus18a.html>)
5. Reuben Binns, Michael Veale, Max Van Kleek and Nigel Shadbolt, 'Like trainer, like bot? Inheritance of bias in algorithmic content moderation' in Giovanni Luca Ciampaglia, Afra Mashhadi and Taha Yasseri (eds), *Social Informatics: 9th International Conference, SocInfo 2017, Proceedings, Part II* (Springer 2017) DOI: 10/cvc2

-
6. Michael Veale, Lilian Edwards, David Eyers, Tristan Henderson, Christopher Millard and Barbara Staudt Lerner, 'Automating Data Rights' in David Eyers, Christopher Millard, Margo Seltzer and Jatinder Singh (eds), *Towards Accountable Systems (Dagstuhl Seminar 18181)* (Dagstuhl Reports 8(4), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik 2018) DOI: 10/gffngz

Abstract

Personal data is increasingly used to augment decision-making and to build digital services, often through machine learning technologies, model-building tools which recognise and operationalise patterns in datasets. Researchers, regulators and civil society have expressed concern around how machine learning might create or reinforce social challenges, such as discrimination, or create new opacities difficult to scrutinise or challenge. This thesis examines how of machine learning systems that matter—those involved in high-stakes decision-making—are and should be governed, in their technical, legal and social contexts.

First, it unpacks the provisions and framework of European data protection law in relation to these social concerns and machine learning's technical characteristics. In chapter 2, how data protection and machine learning relate is presented and examined, revealing practical weaknesses and inconsistencies. In chapter 3, characteristics of machine learning that might further stress data protection law are highlighted. The framework's implicit assumptions and resultant tensions are examined through three lenses. These stresses bring policy opportunities amidst challenges, such as the chance to make clearer trade-offs and expand the collective dimension of data protection rights.

The thesis then pivots to the social dimensions of machine learning on-the-ground. Chapter 4 reports upon interviews with 27 machine learning practitioners in the public sector about how they cope with value-laden choices today, unearthing a range of tensions between practical challenges and those imagined by the 'fairness, accountability and transparency' literature in computer science. One tension between fairness and privacy is unpacked and examined in further detail in chapter 5 to demonstrate the kind of change in method and approach that might be needed to grapple with the findings of the thesis.

The thesis concludes by synthesising the findings of the previous chapters, and outlines policy recommendations going forward of relevance to a range of interested parties.

Impact Statement

Research in this thesis has been strongly motivated by a desire to create *actionable knowledge*. The research questions and approaches used attempt to be responsive to real-world challenges, and the research presented is directly usable by a wide range of actors. Legislators, regulators and activists can draw upon chapters 2 and 3 when looking to enforce, amend or reform data law to cope with emerging challenges. Computer scientists can draw upon chapters 4 and 5 to build systems with methods of tackling social challenges with assumptions that map better to on-the-ground realities. Recommendations for action by these actors are presented in the concluding chapter, chapter 6.

Work deriving from this thesis has arguably already had tangible real world impact, and continues to do so. The work that chapter 2 adapts was described publicly by the UK's Information Commissioner's Office as 'important to the development of [their] thinking',¹ and has been cited by the European Data Protection Board's (in their official regulatory guidance on machine learning),² the Council of Europe in at least three reports,³ the European Commission,⁴ the United Nations,⁵ the German Expert Council for Consumer Affairs,⁶ and RUSI.⁷ This work also formed the basis of several amendments and debates during the passage of the UK's Data Protection Act 2018 through the House of Lords,⁸ something that was noted in the media at the time,⁹ has

¹ Carl Wiper, Information Commissioner's Office, *Presentation to the scientific meeting 'The growing ubiquity of algorithms in society: implications, impacts and innovations'* (The Royal Society, 30 October 2018). Audio file: <https://perma.cc/BWN7-EZQU> (at 27 minutes).

² Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, wp251rev.01* (2018).

³ Committee of experts on internet intermediaries (MSI-NET), *Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications (MSI-NET(2016)06 rev3 FINAL)* (, Council of Europe 2017) (<https://perma.cc/GS4B-ZYHA>); Committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT), *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework (MSI-AUT(2018)05)* (, Council of Europe 2018); Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, *Artificial Intelligence and Data Protection: Challenges and Possible Remedies (T-PD(2018)09Rev)* (, Council of Europe 2018).

⁴ European Commission, *Artificial Intelligence—a European perspective* (2018) DOI: 10 . 2760 / 11251; European Commission, *Automated Decision-Making on the Basis of Personal Data That Has Been Transferred from the EU to Companies Certified under the EU-U.S. Privacy Shield: Fact-Finding and Assessment of Safeguards Provided by U.S. Law* (2018).

⁵ United Nations Special Rapporteur on Extreme Poverty, *Statement on Visit to the United Kingdom, by Professor Philip Alston, United Nations Special Rapporteur on extreme poverty and human rights* (2018) (https://www.ohchr.org/Documents/Issues/Poverty/EOM_GB_16Nov2018.pdf).

⁶ German Expert Council on Consumer Affairs, *Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren* (2018).

⁷ Royal United Services Institute for Defence and Security Studies (RUSI), *Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges* (, RUSI 2018).

⁸ HL Deb 13th November 2017, vol 785, 1862; HL Deb 13th December 2017, vol 787, 1575.

⁹ Rebecca Hill, 'Algorithms, Henry VIII powers, dodgy 1-man-firms: Reg strokes claw over Data Protection

been quoted by US FTC Commissioner Noah Phillips,¹⁰ and was one of five papers that received one of the 9th Privacy Papers for Policymakers Awards at the US Congress to highlight that year's 'must-read' privacy scholarship for US Senators and Congress(wo)men.¹¹ Work from section 3.1 was heavily drawn upon in a report funded by the Information Commissioner's Office¹² and profiled in *The Times*,¹³ work from section 3.2 was profiled by a leading service design company,¹⁴ work from chapter 4 has been explicitly replicated in a private sector context by researchers from Microsoft Research,¹⁵ and work from chapter 5 cited by the UK Government in their *Data Ethics Framework*.¹⁶

Acknowledgements

Writing the research that comprises this PhD has been an intense time, with amazing people.

Lilian Edwards not only brought me into the field of Internet law, but she showed me that it is the best field to be in, ever. I hope I can help preserve this thriving, fun and practically-useful academic sanctuary (largely co-located in countless British pubs, or just on Twitter)—so that it may benefit others as it has benefitted me. Max Van Kleek, whose speed, knowledge and generosity never failed to amaze, brought me into the field of HCI, and his friendship led me to me navigate its quirks and meanders. Reuben Binns has almost always usually asked all the important questions before anyone else has thought of them, and is the most inspiring rolemodel for any interdisciplinary academic I can think of. Jef Ausloos, Frederike Kaltheuner and Seda Gürses showed me how to be rigorous and critical whilst seeking to change the world. Chris Marsden brought his impressive and somewhat paradoxical ability to cheer everyone up and to educate them with dismal, amusing anecdotes of besieged or blundering regulators.

Bill' (*The Register*, 30th October 2017) (<https://perma.cc/A6U2-N9VU>).

¹⁰ Noah Phillips, 'Opening Keynote of Commissioner Noah Joshua Phillips, Washington, DC' (*US Federal Trade Commission*, 6th February 2019) (https://www.ftc.gov/system/files/documents/public_statements/1452828/philips_-_fpf_opening_keynote_2-6-19.pdf).

¹¹ <https://fpf.org/9th-annual-privacy-papers-for-policymakers/>.

¹² Javier Ruiz and Ed Johnson-Williams, *Debates, awareness, and projects about GDPR and data protection: Interim Report for the Information Commissioner's Office for the project: "Making new privacy rights protect and enable people's financial futures"* (Open Rights Group and Projects by If 2018).

¹³ Mark Bridge, 'Siri users are denied access to their data' (*The Times*).

¹⁴ Felix Fischer, 'Alexa, tell me my secrets' (*Projects by If*, 28th November 2018) (<https://www.projectsbyif.com/blog/alexa-tell-me-my-secrets>).

¹⁵ Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miroslav Dudik and Hanna Wallach, 'From Audit to Action: Design Needs for Fairness Monitoring and Decision Support in Machine Learning Product Teams' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM 2019) DOI: 10.1145/3290605.3300830.

¹⁶ Department for Digital, Culture, Media & Sport, *Data Ethics Framework* (HM Government 2018).

Many interdisciplinary fields are based on people finding and building homes in hostile conditions. My supervisor towards the end of this PhD, Arthur Petersen, has been applying different disciplinary perspectives on modelling, in policy context, for a long time, and this work has been very inspiring to me. My secondary supervisor, Anthony Finkelstein, has also been generous with his time in busy periods. Those integrating computer science and law are specialists and building insightful collaborations against the odds. A motley crew of ‘fake lawyers’ have been very important role models. Tristan Henderson’s #loveGDPR meme-skills made surprisingly important and communicable theoretical contributions (consider this a LinkedIn endorsement). Ian Brown’s friendship and support has been invaluable. Both Jat Singh and Josh Kroll have been inspiring with their work and cross-disciplinary ethos. The real lawyers in this space have provided many fun and enlightening times too, including Edina Harbinja, Nora Ní Loideain, Andrew Selbst, Malavika Jayaram, Sandra Wachter, Sylvie Delacroix, Karen Yeung, David Robinson, Marion Oswald, Nadya Purtova, Damian Clifford, Laurens Naudts, Frederik Zuiderveen Borgesius, Jennifer Cobbe, Heleen Janssen, Judith Rauhofer, Nico Zingales, Andres Guadamuz and Chris Millard. Special mention goes to Mireille Hildebrandt who has pushed me, as she has many others, to think about issues in ways I had not before and been extremely generous with kindness, time and support. There’s really way too many of this warm community to list here, and I’m grateful to all the people who I’ve been lucky enough to share ideas and drinks with over the last four years.

Michele Acuto taught me how academics can and should support those around them, and Kira Matus showed me since the very first day of my undergraduate degree that the complex, value-laden problems are the ones worth being a researcher for. I hope I can pay both of your kindness and time forward. Thank you to some people who may not realise they were pivotal: Ewa Luger, Adrian Mackenzie, Harry Armstrong and Lydia Nicholas all invited me to speak or engage with cutting edge-conversations right at the start of my PhD when I had very little to my name, and gave me a foothold in the people and debates as they were emerging and growing. Thanks for taking a risk—I also promise to try and pay it forward. Ine Steenmans and Rocío Carrero were the most amazing and inspiring co-instructors I could have, and showed me how to put heart into teaching. I have had great colleagues especially in Jenny McArthur, Loretta von der Tann, Adam Cooper, Ellie Cosgrave, Irina Brass and Leonie Tanczer among others, and special thanks to my PhD cohort, in particular Enora Robin (without whom I would have waited a lot longer to hear about the ‘algorithmic treatment’ provision in French law), Jeremy Webb, Armela Dino, Cristina Romanelli, Andy Kopp, Lucas Somavilla Croxatto, Emilia Smeds and Zoë Henderson. The bedrock of the

department, the professional services staff, have been fantastic, calling out in particular Steve Morrison, Kelly Lawless, Raphaëlle Moor, Joe Dally-Fitzsimons, Sam King, Rob Ebsworth, Alan Seatwo, and Ruth Dollard. Jason Blackstock was generous with his time during the first part of my PhD.

Thanks also go to the examiners of this thesis, Jack Stilgoe and Kieron O'Hara, for spending the time with my work, and later with me in the defence.

Others generous with their time who I have had inspiring interactions with include Solon Barocas, Bettina Berendt, Jon Crowcroft, Jérémy Grosman, Tyler Reigeluth, Kate Crawford, Meredith Whittaker, Jonnie Penn, Linnet Taylor, Alison Powell, Alison Harcourt, Pablo Suarez, Vasilios Mavroudis, Stephanie Ballard, Adrian Weller, James Lee, Midas Nouwens, Jun Zhao, Johannes Welbl, Sofia Olhede, Margot Kaminski, Timandra Harkness, Zeynep Engin, Sarah Gates, Diana Vlad-Calcić, René Mahieu, and Olivia Stevenson.

This thesis also could not have been undertaken in the way it was without Twitter and the tech policy inhabitants, many of which have already been mentioned but many who haven't. This type of digital public sphere, assessing problems, raising issues and fostering collaborations in the open and across borders, is something on the Internet worth staying up (and fighting) for. Even at the moments when you just want to shout 'no more reports! No more regulatory proposals!' and hide.

The world outside tech policy seems to have shrunk since starting this PhD, but I want to also give thanks to very special people who stood (at least sometimes) outside of this space—Felix Lettau, Natacha Faullimmel, and particularly to Aidan Hermans.

Underpinning all of this has been my family, especially my parents and brother, who have been very supporting of me during this quite intense PhD period. Thank you especially for your love and support in bringing me to this point over the years, which has been invaluable. I'm glad that at least some of this work was translated by journalists so that you can have a little idea of what it is that I seemingly do.

This thesis was funded by the the Engineering and Physical Sciences Research Council (EPSRC), under grant number EP/M507970/1. Other funding that supported work done in parallel to this thesis came from UCL Public Policy, UCL STEaPP's 'pump priming' fund, and from the Global Facility for Disaster Reduction and Recovery (GFDRR).

Contents

Table of UK Cases	14
Table of UK Legislation	14
Table of EU Legislation	15
Table of EU Cases	17
Table of Parliamentary Material and Draft Legislation	18
I. Introduction	23
1. Hello, World!	25
1.1. On Algorithms	27
1.2. Machines that Learn	32
1.3. Making the Machines that Learn	35
1.4. Algorithmic War-Stories	39
1.4.1. Recidivism and Racism	40
1.4.2. Discrimination and Search Engines	42
1.4.3. Know Your Customers	43
1.4.4. Bias in the Toolkit	45
1.4.5. Was It Facebook Wot Won It?	47
1.5. Algorithmic issues	48
1.6. Computing to the Rescue?	52
1.6.1. Explaining	54
1.6.2. Debiasing	65
1.6.3. Accounting	71
1.7. Research Questions and Method	73

II. Data Protection Law	81
2. The Law of Machine Learning?	83
2.1. A Regulatory Mixtape	85
2.2. Data Protection Rights and Machine Learning	91
2.2.1. Automated decision prohibitions	91
2.2.1.1. Applicability	96
2.2.1.2. The nature of ‘decisions’	101
2.2.2. Information and Explanation Rights	106
2.2.2.1. Access rights	107
2.2.2.2. Automated decision transparency	110
2.2.2.2.1. Article 22	111
2.2.2.2.2. Articles 13–15	112
2.3. Problems of data protection individualism	116
2.3.1. The transparency fallacy	116
2.3.2. Weak collective provisions	119
2.4. Beyond individualism	128
2.4.1. Purpose limitation	128
2.4.2. Data Protection Impact Assessments	131
2.4.3. Certification	134
3. Data Protection’s Lines, Blurred by Machine Learning	137
3.1. Line 1: People from Data	139
3.1.1. Introduction	140
3.1.2. Background	143
3.1.3. Vignette of Vanished Rights 1: TfL Wi-Fi	148
3.1.4. Vignette of Vanished Rights 2: Apple’s ‘Siri’ voice assistant	154
3.1.5. Rescuing Data Protection by Design	158
3.1.5.1. Parallel Systems for Data Rights	158
3.1.5.1.1. Obligations to Retain Data	159
3.1.5.1.2. Acquiring Additional Information	162
3.1.5.2. Making Trade-offs Accountable	164
3.1.5.3. Information Rights around Privacy Architectures	166
3.1.6. Interim remarks	172
3.2. Line 2: Data from Models	172
3.2.1. Why Control Models?	173
3.2.2. Models on the Move	175
3.2.3. Inverting Models	176

3.2.4. Models as Personal Data?	180
3.2.5. Implications for Data Subjects	183
3.2.5.1. Information Rights	183
3.2.5.2. Erasure Rights	185
3.2.5.3. Restriction and Objection Rights	187
3.2.6. Implications for Data Controllers	188
3.2.6.1. Security Principle	188
3.2.6.2. Storage Limitation	191
3.2.7. Interim discussion	191
3.3. Line 3: Data from Sensitive Data	192
3.3.1. Automated Lipreading	193
3.3.2. Trajectories	194
3.3.3. Applications and Concerns	196
3.3.4. Limitations	200
3.3.5. Regulatory challenge	202
3.3.5.1. Crime detection uses	207
3.3.5.2. Non-crime detection uses	208
3.3.5.3. The (im)possibility of anonymisation	209
3.3.6. Privacy by design as an escape route?	211
3.3.6.1. Data minimisation	212
3.3.6.1.1. Sensitive data minimisation	214
3.3.6.1.2. Anonymisation	215
3.3.6.2. Purpose limitation	216
3.3.6.3. Objection	217
3.3.7. Interim discussion	219
3.4. Summary remarks	220

III. Machine Learning on the Ground 221

4. Coping with Value(s) in Public Sector Machine Learning 223	223
4.1. Automation Systems	224
4.2. Augmentation Systems	226
4.3. Public Values	229
4.4. Questions	232
4.5. Method	233

4.6. Findings	236
4.6.1. Internal actors and high-stakes machine learning	236
4.6.1.1. Getting individual and organisational buy-in	237
4.6.1.2. Over-reliance, under-reliance and discretion	238
4.6.1.3. Augmenting models with additional knowledge	240
4.6.1.4. Gaming by decision-support users	241
4.6.2. External actors and high-stakes machine learning	242
4.6.2.1. Sharing models and pushing practices	242
4.6.2.2. Accountability to decision subjects	243
4.6.2.3. Discriminating between decision subjects	243
4.6.2.4. Gaming by decision subjects	245
4.7. Implications for research and practice	246
4.7.1. ‘The probability of being correct tanked’: Data changes	246
4.7.2. ‘Always a person involved’: Augmenting outputs	248
4.7.3. ‘When it aligns with intuition’: Understanding discretion	249
4.7.4. ‘I’m called the single point of failure’: Moving practices	251
4.7.5. ‘Looks like we’ve 100% accuracy’: Talking performance	252
4.8. Interim conclusions	253
5. Unpacking a tension: ‘Debiasing’, privately	255
5.1. Knowing protected characteristics: necessary but problematic	255
5.2. ‘Debiasing’ with limited sensitive data	257
5.2.1. Trusted third parties	258
5.2.1.1. As ex post discrimination detector	260
5.2.1.2. As ex ante discrimination mitigator	261
5.2.1.3. Who could be a third party?	263
5.2.1.4. Cryptographic ‘third parties’	265
5.2.2. Fairness knowledge bases	265
5.2.2.1. Practical considerations	267
5.2.2.2. Confounders	270
5.2.3. Exploratory fairness analysis	271
5.2.3.1. With unsupervised learning	272
5.2.3.2. With interpretable models	273
5.2.4. Discussion	274

IV. Joining the Dots	279
6. Synthesis and Recommendations	281
6.1. Going beyond canonical ‘decisions’	284
6.2. Building and scrutinising process	285
6.3. Anticipation needs	286
6.4. Rethinking the role of groups	288
V. In the Back	293
7. Appendices	295
7.1. Apple Correspondence	295
7.2. Workshops and Conferences Attended	301
Bibliography	307

List of Figures

1.1. The machine learning pipeline	35
1.2. Debiasing approaches in the context of the machine learning pipeline	67
2.1. Articles 7 and 8, Charter of Fundamental Rights	86
3.1. Lines drawn by data protection and blurred by machine learning	138
3.2. Model inversion and membership inference attacks.	178
3.3. Overfitting	186
5.1. Three approaches to ‘debiasing’ without holding sensitive characteristics.	258

List of Tables

4.1. Public sector values relevant to machine learning.	230
---	-----

Table of UK Cases

Durant v Financial Services Authority [2003] EWCA Civ 1746 ...	204	R (Moseley) v London Borough of Haringey [2014] UKSC 56	133
Edem v Information Commissioner [2014] EWCA Civ 92	204	R v Luttrell (Gerrard Francis) [2004] EWCA Crim 1344 ...	203, 207
Northern Metco Estates Ltd v Perth and Kinross DC 1993 SLT (Lands Tr) 28	124	Southampton City Council v Information Commissioner [2013] UKFTT 20120171 (GRC)	211

Table of UK Legislation

Data Protection Act 2018	87	s 35	
s 14	95	(2)	207
s 30	208	(4)	207

(5)	207	2000	
s 49	95	s 234C	121
s 188	121	Freedom of Information Act 2000	
s 189	121	s 51	125
part 3	207	Offensive Behaviour at Football and	
sch 1 para 8	257	Threatening Communications	
Enterprise Act 2002		(Scotland) Act 2012	208
s 11	121	The Enterprise Act 2002	
Equality Act 2010		(Super-complaints to	
s 4	65	Regulators) Order 2003, SI	
Financial Services (Banking Reform)		2003/1368	121
Act 2013		The Equality Act 2010 (Gender Pay	
s 68	121	Gap Information) Regulations	
Financial Services and Markets Act		2017, SI 2017/172	126

Table of EU Legislation

Reg 2012/C [2012] OJ C326/391		art 12	109, 159, 169
art 7	85	art 15	93
art 8	70, 85, 90, 167	art 17	144
(2)	164	art 30	
art 11	167	(1)	114
art 21	70, 167	(3)	114
Dir 2016/680 [2016] OJ L119/89 ..	87, 207	recital 46	144
art 10	207	Dir 96/9/EC [1996] OJ L77/20 .	83, 173
art 11	95	Dir 2002/58/EC [2002] OJ L201/37	87
art 20	145	Reg 2016/679 [2016] OJ L119/1 ...	87, 169
Dir 2016/943 [2016] OJ L157/1		art 2	
recital 35	170	(1)	173, 187, 207
Dir 95/46/EC [1995] OJ L281/31		(2)	207
		art 4	

TABLE OF EU LEGISLATION

(1)	88, 180, 203	art 16	89, 109
(2)	88, 187	art 17	89, 109, 116, 147, 185
(4)	106	art 18	89, 109, 187
(5)	139, 180, 197, 210	art 20	142, 147
(7)	88	art 21	89, 109, 147, 152, 187, 217
(11)	118	(5)	123
art 5		art 22	112, 168, 188
(1)	88, 143, 146, 216	(1)	126
(2)	88, 146, 162, 170	(2)	94, 99
(c)	212, 213	(3)	95
(d)	205	(4)	94, 127
art 6	94	art 24	
(1)	89, 206	(1)	170
(4)	89	(a)	113
(e)	206	art 25	90, 145
(f)	206	art 35 ...	89, 90, 95, 112, 131, 164
art 7	88	(1)	100
(1)	118	(3)	100
(2)	118	(5)	132
(4)	118, 130	(7)	165
art 9	94, 158, 214	(9)	133
(1)	89, 206, 217, 256	art 36	90, 134, 171
(2)	127, 206	(3)	170
art 11	140	art 37	90
(1)	184	art 38	
(2)	140	(3)	90
art 12		art 40	90
(4)	167	art 42	90, 164, 217
art 13	167	(1)	134
(1)	89, 184	(4)	135
(2)	95, 100, 112, 166, 167	(6)	135
art 14	167	(8)	135
(2)	100, 112, 166, 167	art 49	94
(5)	184	art 58	
art 15	107, 147	(1)	124
(1)	100, 112, 184	art 70	
(3)	89	(1)	114

art 80		recital 63	170
(1)	120, 171	recital 69	152
(2)	121, 171	recital 71	71, 105, 112, 167
art 83		recital 75	165
(2)	135	recital 84	259
recital 26	140, 163, 173, 181, 205, 206, 210, 214	Reg 2018/1807 [2018] OJ L303/59	
recital 27	120	art 2	
recital 28	139, 214	(1)	173
recital 29	214	Dir 765/2008 [2018] OJ L218/30	135
recital 30	157	Treaty of Lisbon, 2007/C [2007] OJ	
recital 32	210	C306/1	85
recital 43	210	Treaty on the Functioning of the	
recital 47	206	European Union [2016] OJ	
recital 51	214, 215	C202/1	
recital 57	163	art 5	
recital 59	167	(3)	103
		art 16	85

Table of EU Cases

Association Belge des Consommateurs
Test-Achats ASBL and Others
v Conseil des ministres
(C-236/09)
ECLI:EU:C:2011:100 50

College van burgemeester en
wethouders van Rotterdam v
MEE Rijkeboer (C-553/07)
ECLI:EU:C:2009:293 109,
159, 160

Fashion ID GmbH & Co KG v
Verbraucherzentrale NRW eV
(joined parties: Facebook
Ireland Limited,
Landesbeauftragte für
Datenschutz und
Informationsfreiheit
Nordrhein-Westfalen)
(Opinion of AG Bobek
C-40/17)
ECLI:EU:C:2018:1039 ... 106,

119	
František Ryneš v Úřad pro ochranu osobních údajů (C-212/13) ECLI:EU:C:2014:2428 ...	207
Google Spain v Agencia Española de Protección de Datos (AEPD) and González (C-131/12) ECLI:EU:C:2014:317	105, 115, 116, 287
Maximillian Schrems v Data Protection Commissioner (C-362/14) ECLI:EU:C:2015:650	110
Patrick Breyer v Bundesrepublik Deutschland (C-582/14) ECLI:EU:C:2016:779	181, 182, 205
Peter Nowak v Data Protection Commissioner (C-434/16) ECLI:EU:C:2017:994	109, 110, 181, 204, 205
Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v Wirtschaftsakademie Schleswig-Holstein GmbH (C210/16) ECLI:EU:C:2018:388	105, 287
YS v Minister voor Immigratie, Integratie en Asiel and Minister voor Immigratie, Integratie en Asiel v M and S (C-141/12) ECLI:EU:C:2014:2081 ...	182, 204

Table of Parliamentary Material and Draft Legislation

Freedom of Information (Extension) Bill HC Bill (2017–19) [23]	125
AI in the UK: ready, willing and able? (Select Committee on Artificial Intelligence (Lords), HL 2018, Paper)	59
Algorithms in Decision-Making (Science and Technology Committee (Commons), HC 2018, 351)	59
Official Report, House of Commons 13th March 2018 vol 637	104

TABLE OF PARLIAMENTARY MATERIAL AND DRAFT LEGISLATION

Official Report, House of Lords	vol 787	5, 115
vol 785	5, 115	
	vol 788	121, 122

Acronyms

A29WP Article 29 Data Protection Working Party. 81, 83, 84, 87, 88, 100, 101, 104, 113, 114, 130, 136, 147, 149–151, 163, 165, 183, 193

AG Advocate General. 92, 104

AI artificial intelligence. 19, 71, 84, 226, 241

ALRS automated lipreading system. 174, 176–178, 182–184, 187, 189, 190, 193, 195, 263

API application programming interface. 125, 178, 238, 242

AVARD Anonymous Video Analytics for Retail and Digital Signage. 195, 196

CCTV closed-circuit television. 123, 173, 176, 177, 182, 186–188, 190, 196

CFAA Computer Fraud and Abuse Act. 110

CJEU Court of Justice of the European Union. 84, 85, 91, 92, 95, 96, 98, 100, 101, 104, 119, 124, 141, 142, 152, 162, 163, 183, 263

CMA Competition and Markets Authority. 241

CNIL Commission nationale de l’informatique et des libertés. 79, 80, 91

COMPAS Correctional Offender Management Profiles for Alternative Sanctions. 32, 33, 45

DADM discrimination-aware data mining. 57, 237, 240, 243, 244, 246

DARPA the Defense Advanced Research Projects Agency. 45, 46

DCMS the Department for Digital, Culture, Media and Sport. 241

DNT Do Not Track. 108, 136

DPA data protection authority. 77, 83, 103, 106, 107, 109, 113, 118, 119, 127, 132, 135, 152, 153, 165, 182

- DPbD** data protection by design. 128, 131, 141, 144, 147, 149–153, 174, 191, 263
- DPD** Data Protection Directive 1995. 80, 81, 91, 95, 96, 100, 105, 126, 127, 141, 142, 152, 162, 164, 173
- DPIA** data protection impact assessment. 82, 86, 116–118, 120, 131, 132, 147, 148, 152, 155, 236, 261, 262, 264
- DPO** data protection officer. 77
- ECHR** European Convention on Human Rights. 190
- ECOA** Equal Credit Opportunity Act. 48
- EDPB** European Data Protection Board. 84, 100, 119, 183
- EDPS** European Data Protection Supervisor. 107, 114, 135
- EFF** Electronic Frontier Foundation. 115
- EPA** Environmental Protection Agency. 247
- EPSRC** the Engineering and Physical Sciences Research Council. 6, 45
- FATML** ‘fairness, accountability and transparency in machine learning’. 57, 60–62, 209, 213, 223, 237, 240, 243, 244, 246
- FCRA** US Fair Credit Reporting Act. 48
- FIP** Fair Information Practice. 128, 130
- FOI** freedom of information. 93, 105, 110, 118, 132, 152, 153, 261
- FPF** Future of Privacy Forum. 136
- GAFA** Google, Apple, Facebook and Amazon. 19, 104
- GDPR** General Data Protection Regulation 2016. 59, 63, 74, 75, 77, 78, 80, 82, 84, 86, 87, 91, 93, 95–98, 102, 103, 105–108, 111–116, 118–124, 126, 128, 130, 131, 135, 140, 142–144, 146, 147, 151–153, 155, 161, 162, 164, 165, 172, 183–187, 196, 198, 202, 234, 236, 257, 258, 261, 262, 265
- HCI** human–computer interaction. 20, 48, 229, 231
- ICAAIL** the International Conference of AI and Law. 71

ICDPPC International Conference of Data Protection and Privacy Commissioners. 79

ICML the International Conference on Machine Learning. 45

ICO Information Commissioner's Office. 77, 83, 101, 110, 116, 125, 132, 133, 135, 177, 190, 241

IP intellectual property. 247

JURIX the International Conference on Legal Knowledge and Information Systems. 71

MAC media access control. 132–136

NeurIPS Neural Information Processing Systems. 45

PbD privacy by design. 116, 127, 128, 130, 136, 138, 139, 154, 190, 198

PCA principal component analysis. 249

PET privacy-enhancing technology. 126, 127, 129, 130, 134, 142, 154

PIA privacy impact assessment. 116

SAC Sustainable Apparel Coalition. 247

SAR subject access request. 76, 93, 183

SPFL Scottish Professional Football League. 187

SSI silent speech interface. 176

TfL Transport for London. 121, 132–135, 153, 188

XAI explainable artificial intelligence. 45

Part I.

Introduction

1. Hello, World!

Competing narratives underpinning the opportunities and risks of ‘data-driven’ decision-making will be more than a footnote in history books looking back on these decades. One possible narrative might read as follows.

Data accumulation and analysis demonstrated its potential in the business models of the late 20th and early 21st centuries. The economic and political success of technology giants such as Google, Apple, Facebook and Amazon (GAFA), who swelled to be among the most valuable companies in the world through operationalising unprecedented amounts of structured data, excited governments and businesses in other sectors—particularly in relation to the potential utility of data collection and retention for their own purposes. Seeking to mimic the practices that appeared to underpin GAFA’s rise, they internalised a new ‘logic of accumulation’, and saw these firms as ‘emissaries of the future’.¹⁷ Attendees of presentations on the area in corporate domains and more widely were commonly told that the current wave of excitement results from ‘technical developments in the field, increased availability of data, and increased computing power.’¹⁸

Still, even the largest actors now accumulating (or simply appreciating) datasets often struggled to see how these could be transformed into the sheer value as GAFA had. Sizeable datasets, computing and machine learning technologies had already been relatively mature for some decades: the allure of ‘Big Data’ was just as much a social, cultural and mythological phenomenon as a technical one.¹⁹ To keep the hope (and hype) alive, industry interest grew in modern analytics rather than the datasets: on the models and insights possible to extract from these information pools. Even though they were far from practical problems or application areas, high-profile research successes such as Google subsidiary Deepmind’s ‘Alpha Go’ system’s 2016 victory against high-ranking professional Go player Lee Sedol²⁰ made it seem like machine learning

¹⁷ Shoshana Zuboff, ‘Big other: surveillance capitalism and the prospects of an information civilization’ (2015) 30(1) *Journal of Information Technology* 75 DOI: 10/gddxpv, 85.

¹⁸ See eg The Royal Society, *Machine learning: The power and promise of computers that learn by example* (The Royal Society 2017) 18.

¹⁹ danah boyd and Kate Crawford, ‘Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon’ (2012) 15(5) *Information, Communication & Society* 662 DOI: 10/7vq.

²⁰ Elizabeth Gibney, ‘Google AI Algorithm Masters Ancient Game of Go’ (2016) 529(7587) *Nature* 445 DOI: 10/bb5s.

1. Hello, World!

(increasingly branded synonymously with ‘artificial intelligence’, or AI) could do just about anything. This in turn thrilled governments who saw new sectors and possibilities for economic growth amidst stagnating traditional industries and in the aftermath of the 2008 financial crisis.²¹

A range of scholars from across domains, building on wide existing literatures including social sorting and surveillance, science and technology studies, histories of statistics and modelling, human geography, technology law and expert systems, all warned in their own ways that all was not as value-free and universally positive as it seemed. The core concern was that these systems and problem-solving approaches would enter use veiled in ‘fake’ neutrality and objectivity, and both mask and exacerbate ongoing issues of power, inequality, and challenges to a range of fundamental rights and freedoms.²²

Surprising some, these objections did *not* fall totally on deaf ears, but instead collected a substantial amount of high-profile interest in a range of domains, including computer science, statistics and law. Machine learning conferences included tracks considering aspects of the social impacts of technologies being presented in just the next room along. Data protection law was driven from a minor area of practice in the dingiest corner of law firms’ offices to headlines around the world. Tech giant IBM even aired an advert proclaiming their desire to fight bias in artificial intelligence during the 2019 Academy Awards—some of the most expensive television estate money can buy. The appropriate skillset and role of practitioners of data analysis, increasingly referred to as ‘data scientists’ has also been central to debate, with codes of conduct, certification procedures, oaths, diversity requirements, unionised action, or ethical training programmes all proposed.

This thesis focuses on the governance of machine learning technologies with consequential outcomes, with a (normative) focus on how in practice we might avoid, or at least steer and manage, many of the ‘algorithmic harms’ identified in these debates and discussions. Three distinct perspectives are combined throughout this thesis: technology law and policy; human–computer interaction; and computer science. Chapters 2 and 3 primarily draw upon data protection law in the context of computer science research, while chapters 4 and 5 draw primarily upon human–computer interaction (HCI), drawn broadly, again in technical context. More detailed structure

²¹ See eg Wendy Hall and Jérôme Pesenti, *Growing the artificial intelligence industry in the UK* (HM Government 2017) (<https://perma.cc/3E45-MYSM>) (the Hall-Pesenti review of AI for the UK government) and State Council of the People’s Republic of China, *A Next Generation Artificial Intelligence Development Plan* (Rogier Creemers, Graham Webster, Paul Tsai and Elsa Kania trs, Government of China 2017) (<https://perma.cc/9EE3-4MXH>) (China’s 2017 AI Strategy).

²² See eg Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker and Kate Crawford, *The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term* (AI Now Institute 2017) (<https://perma.cc/G9AX-XQFN>) (a report summarising many issues and communities).

for this thesis going forwards is given towards the end of this chapter.²³

Firstly, I will explore and expand upon the topic of this thesis in the context of contemporary literature and debates.

1.1. On Algorithms

‘Algorithms’ are far from new when defined according to their classic meaning in the formal sciences. The term covers recipes, rules, procedures, techniques, processes, procedures and methods, and over time has come to mean ‘any process that can be carried out automatically’.²⁴ For a detailed computer science definition, there may be few places more apt than the highly popular textbook *Introduction to Algorithms*. Despite its 1,000 pages, it sold its 500,000th copy in 2011.²⁵ The text starts with the following two perspectives on the nature of algorithms. The first is as follows:

Informally, an algorithm is any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output. An algorithm is thus a sequence of computational steps that transform the input into the output.²⁶

This flavour of definition, which I refer to as a *repeatable recipe* definition, is commonly cited at the beginning of work on the governance of algorithms or machine learning systems.²⁷ At interdisciplinary workshops or conferences covering the approaches to harms I will go on to discuss,²⁸ I have heard many frustrated computer scientists call for their social science colleagues to align their use of the word ‘algorithm’ with a definition like this.²⁹ Some social science scholars have even argued that there is a moral imperative to ensure that definitional alignment exists between engineers

²³ See section 1.7.

²⁴ Jean-Luc Chabert, Évelyne Barbin, Jacques Borowczyk, Michel Guillemot, Anne Michel-Pajus, Ahmed Djebbar and Jean-Claude Martzloff, *A history of algorithms: From the pebble to the microchip* (Chris Weeks tr, Springer 1999) 2.

²⁵ Larry Hardesty, ‘Milestone for MIT Press’s bestseller’ (*MIT News*, 10th August 2011) (<http://news.mit.edu/2011/introduction-to-algorithms-500k-0810>).

²⁶ Thomas H Cormen, Charles E Leiserson, Ronald L Rivest and Clifford Stein, *Introduction to algorithms* (MIT Press 2009) 5.

²⁷ See eg Mike Ananny, ‘Toward an ethics of algorithms: Convening, observation, probability, and timeliness’ (2016) 4(1) *Science, Technology & Human Values* 93 DOI: 10/gddv77; Rob Kitchin, ‘Thinking critically about and researching algorithms’ (2017) 20(1) *Information, Communication & Society* 14 DOI: 10/gc3hsj; Lucas Introna and David Murakami Wood, ‘Picturing algorithmic surveillance: The politics of facial recognition systems’ (2002) 2(2/3) *Surveillance & Society* DOI: 10/gdxwfx.

²⁸ See eg section 1.6. See also a list of workshops attended during writing this thesis on p. 301.

²⁹ For work tracking this dynamic, see Nick Seaver, ‘Algorithms as Culture: Some tactics for the ethnography of algorithmic systems’ (2017) 4(2) *Big Data & Society* DOI: 10/gd8fdx.

1. Hello, World!

and social science in order to better facilitate the transmission of core social scientific concerns into engineering practice.³⁰

While shared definitions can act as a ‘good enough’ roadmap for discussion across different domains,³¹ they need to be faithful enough to each domain to be useful at traversing the ‘boundary’ between them. Is the *repeatable recipe* capable of such boundary-hopping? In my view, it is not.

Firstly, the *repeatable recipe* approach is enormous in the range of systems it covers. Any well-specified series of instructions meets these criteria. It only has to produce ‘some value’ as output—not a useful one, nor one which relates to any given task. While a machine learning model is an algorithm, so is long division, or the process of determining tax at a restaurant. This is not to say that simple, rote formulae cannot be of consequence—powerful institutions have long deployed them towards certain human purposes to sort and classify people, places, objects and phenomena. Defined as such, these algorithms are old and constant features of civic life—present for centuries, discussed by scholars and activists for decades.³² Max Weber’s idealised bureaucracy, for example, would operate ‘according to calculable rules’ and ‘without regard for persons’.³³ Mechanisms to render populaces countable or legible have been both common and consequential throughout history, and it has been argued their use has often come at great and largely unrecognised human cost.³⁴ Such concerns continue to resound against today’s practices in areas such as official statistics, where critics have illustrated out how problematic and misleading assumptions (such as relating to violence against women) have been integrated into international accounting mechanisms.³⁵

This thesis does not endeavour to consider the social aspects of all classification or quantification. If there *is* something new or interesting about machine learning bey-

³⁰ See Paul Dourish, ‘Algorithms and their others: Algorithmic culture in context’ (2016) 3(2) *Big Data & Society* DOI: 10/gcdx9q, 2.

³¹ Susan Leigh Star and James R Griesemer, ‘Institutional Ecology, Translations and Boundary Objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39’ (1989) 19(3) *Social Studies of Science* 387 DOI: 10/ckpxb6, 410.

³² What scholars now discuss as algorithms have also been discussed under various other labels. These include (non-synonymously) expert systems, cybernetic systems, artificial intelligence, robotic process automation, indicators, autonomous systems, rule-based systems, computational models, risk-scoring systems and analytic models among many other guises. This thesis attempts to link to these literature as and when appropriate throughout.

³³ Max Weber, ‘Bureaucracy’ in H H Gerth and C Wright Mills (trs), *From Max Weber* (Routledge 1958) 215.

³⁴ James C Scott, *Seeing like a state: How certain schemes to improve the human condition have failed* (Yale University Press 1998).

³⁵ For discussions of these topics, see eg Geoffrey Bowker and Susan Leigh Star, *Sorting things out: Classification and its consequences* (The MIT Press 1999); David Lyon, *Surveillance as social sorting: Privacy, risk, and digital discrimination* (Routledge 2003); Richard Rottenburg, Sally Engle Merry, Sung-Joon Park and Johanna Mugler, *The world of indicators: The making of governmental knowledge through quantification* (Cambridge University Press 2015); Sally Engle Merry, *The seductions of quantification: Measuring human rights, gender violence, and sex trafficking* (University of Chicago Press 2016).

and the valuable existing studies of social aspects of sorting, accounting and valuing individuals, it is likely to be difficult to find by starting with a very wide definitional net.

Secondly, it is unclear whether this definition is even the one held by all relevant ‘technical’ actors, let alone across very different disciplines. Boundary objects such as shared schema and definitions do not only need to successfully traverse disciplines of thought, but also different modes of practice. This seems particularly key in uncertain, value-laden areas of science such as machine learning, which are broadly held to require new types of engagement with citizens and society beyond review by academics and researchers.³⁶ As Nick Seaver argues, while it seems relatively self-evident that a ‘data scientist working at Facebook in 2017, a university mathematician working on a proof in 1940, and a doctor establishing treatment procedures in 1995 may all claim, correctly, to be working on “algorithms,” this does not mean they are talking about the same thing.’³⁷ The definitions of the term, or the locations they identify it in or with, are far from stable.³⁸ The parts of the system that different concerned individuals find relevant and important differs, and interacting or observing with this broader system from a wide range of angles can change the boundaries or nature of the ‘algorithm’ itself.

If we are interested in where actors feel the ‘action happens’, we can neither rely on an *a priori* definition of an algorithm to guide us, nor indeed for there to be a stable definition at all. Luckily (and potentially indicative of why it has sold over 500,000 copies), *Introduction to Algorithms* provides a second perspective.

We can also view an algorithm as a tool for solving a well-specified computational problem. The statement of the problem specifies in general terms the desired input/output relationship. The algorithm describes a specific computational procedure for achieving that input/output relationship.³⁹

This definition recognises that like technologies in general, algorithms are directed towards *certain human purposes*. In that sense, it is *teleological* in nature: an explanation of algorithms as a function of their end, purpose or goal. Classic definitions of technology incorporate this notion. Harvey Brooks, for example, described technology as ‘knowledge of how to fulfil certain human purposes in a specifiable and reproducible way’.⁴⁰ This approach to algorithms still emphasises the ‘specific computational pro-

³⁶ Silvio O Funtowicz and Jerome R Ravetz, ‘Science for the Post-Normal Age’ (1993) 25(7) *Futures* 739 DOI: 10/fqntk9 (on the notion of ‘extended peer review’).

³⁷ Seaver, ‘Algorithms as Culture’ (n 29) 2.

³⁸ *ibid.*

³⁹ Cormen, Leiserson, Rivest and Stein (n 26) 5.

⁴⁰ Harvey Brooks, ‘Technology, evolution, and purpose’ (1980) 109(1) *Daedalus* 65, 66.

1. Hello, World!

cedure'. Whether this is the best approach is open to questioning for a range of reasons.

Algorithms have never been divorced from their environment. Where systems that name and count people are considered, they do not always do so passively, but in creating a classification system, have often shaped the people they name in a 'looping effect'.⁴¹ Whether or not an algorithm succeeds in fulfilling the envisaged human purposes, or creating the desired input-output relationship is not only contingent on the quality of the programming, but on the dynamics of the phenomena it is intended to be applied to. In this sense, algorithms are neither technical nor non-technical; neither neutral nor imbued with a single, static set of values.

Perhaps the clearest current example that illustrates that such a 'looping effect' can go beyond names is the current, algorithmically-charged, debate around self-driving vehicles and the future of the city. As researchers discover the difficulties of dealing with a chaotic, pedestrian environment, some high profile individuals at large technology companies have suggested it is not the car, but the city and the pedestrian, that needs to change. Andrew Ng, a prominent machine learning researcher and co-founder of self-driving vehicle firm Drive.AI, argued that while these vehicles were improving at detecting pedestrians, governments and companies should instead, or at least in tandem, 'ask people to be lawful and considerate', as '[s]afety isn't just about the quality of the AI technology'—a proposal which quickly came under heavy fire⁴² (and which echoed the manner through which the automobile lobby reshaped streets and cities in the US in the first half of the 20th century⁴³). Critics also noted that many *existing* road safety problems would benefit from such behaviour change, and that changing behaviour is harder than it would seem.⁴⁴ If these technologies are deployed, pedestrians and other users of public space and infrastructure *are* likely to change, react and adapt, but may not do so as expected.

Compounding this is the changing nature of software development to integrate factors outside what have typically been seen as the boundaries of deployed code. Software engineering practice in general has been migrating from a shrink-wrapped

⁴¹ Ian Hacking calls this 'dynamic nominalism'. See Ian Hacking, 'Kinds of People: Moving Targets' in PJ Marshall (ed), *Proceedings of the British Academy, Volume 151* (British Academy 2007) DOI: 10/gfgrbm.

⁴² For the original see Russell Brandom, 'Self-Driving Cars Are Headed toward an AI Roadblock' (*The Verge*, 3rd July 2018) (<https://www.theverge.com/2018/7/3/17530232/self-driving-ai-winter-full-autonomy-waymo-tesla-uber>). For a variety of responses from academics and practitioners, see Jeremy Kahn, 'To Get Ready for Robot Driving, Some Want to Reprogram Pedestrians' (*Bloomberg*, 16th August 2018) (<https://www.bloomberg.com/news/articles/2018-08-16/to-get-ready-for-robot-driving-some-want-to-reprogram-pedestrians>).

⁴³ See generally Peter D Norton, *Fighting Traffic: The Dawn of the Motor Age in the American City* (MIT Press 2008).

⁴⁴ Kahn (n 42).

model to an ever-changing *agile* approach.⁴⁵ Within this paradigm, the software development process expands to involve constant feedback from and experimentation on users: deploying, monitoring and evaluating of new features in near real-time.⁴⁶ Exacerbating this dynamic software development trend is the use of machine learning components (described below⁴⁷) designed to take new input data and change over time. These components directly couple the users and the phenomena being modelled, and make such an explicit coupling part of the codebase.⁴⁸ There is a growing use of reputation systems in social platforms which can have undesirable emergent properties,⁴⁹ while some applications seek to catalyse and rely on social machines (of which Wikipedia is usually cited as an example), connecting human work with machine work to achieve higher level goals not capable by one of the two alone.⁵⁰ All of these changes have more tightly bound a software object to the environment it is situated within: changes in individual user behaviour and the broader measured phenomena themselves now alter the characteristics of software being produced. Software has never been divorced from its environment, but it is arguably now closer and more tangled than ever. The classic minimal computer programme *Hello, World!* has never seemed so apt.

This multitude of interactions that exist in and around algorithms call for them to be better considered through a *systems lens*. A system is an interconnected set of elements that is coherently organised in a way that achieves something.⁵¹ The notion that a system ‘achieves something’ is familiar from the functional focus of the second ‘algorithm’ definition from *Introduction to Algorithms* above, and from teleological perspectives in general. The coherent organisation too echoes the ‘specific’ nature of that description of an algorithm. The new focus a systems lens provides is *intercon-*

⁴⁵ Seda Gürses and Joris van Hoboken, ‘Privacy after the Agile Turn’ in Evan Selinger, Jules Polonetsky and Omer Tene (eds), *The Cambridge Handbook of Consumer Privacy* (Cambridge University Press 2018) DOI: 10/gfgq84.

⁴⁶ *ibid.*

⁴⁷ See section 1.2.

⁴⁸ See eg Danielle Ensign, Sorelle A Friedler, Scot Neville, Carlos Scheidegger and Suresh Venkatasubramanian, ‘Runaway Feedback Loops in Predictive Policing’ in *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAT*)* (2018) (on the connection between how police use decision-support systems and the development of a dynamic model over time).

⁴⁹ Alex Rosenblat, Karen EC Levy, Solon Barocas and Tim Hwang, ‘Discriminating Tastes: Uber’s Customer Ratings as Vehicles for Workplace Discrimination’ (2017) 9(3) *Policy & Internet* 256 DOI: 10/gddxqn; Jevan A Hutson, Jessie G Taft, Solon Barocas and Karen Levy, ‘Debiasing Desire: Addressing Bias & Discrimination on Intimate Platforms’ (2018) 2(CSCW) *Proc. ACM Hum.-Comput. Interact.* 73:1 DOI: 10/gfkwxxv.

⁵⁰ Jim Hendler and Tim Berners-Lee, ‘From the Semantic Web to social machines: A research challenge for AI on the World Wide Web’ (2010) 174(2) *Artificial Intelligence* 156 DOI: 10/cr4qpw; Nigel R Shadbolt, Daniel A Smith, Elena Simperl, Max Van Kleek, Yang Yang and Wendy Hall, ‘Towards a classification framework for social machines’ in *Proceedings of the 22nd International Conference on World Wide Web* (2013) DOI: 10/gfgq9b.

⁵¹ Donella H Meadows, *Thinking in systems* (Earthscan 2008) 11.

1. Hello, World!

nectedness. What sets apart the algorithms focussed upon in this thesis is how they are interwoven with other actors, software, hardware, and their environments, both dynamically in real-time through the processes surrounding their deployment, and historically through the many actors that already shaped the data used to build, tune and tailor them. They are ‘not standalone little boxes, but massive, networked ones with hundreds of hands reaching into them, tweaking and tuning, swapping out parts and experimenting with new arrangements’.⁵² The swapping and the tweaking might be consciously carried out by developers, or could be a result of actions by individuals or groups who are unaware they are interacting with a system.

This thesis thus does not focus upon *algorithms*, but on *algorithmic systems*.

Incidentally, a systemic approach to algorithms and their wider surroundings, which emphasises the behaviour of properties which emerge from the system, should not be confused with a functionalist approach. Those individuals or organisations designing systems generally have intent, and bring purposes they want to direct their efforts towards, but the congruence of this intent with the function of the system as a whole is far from assured. Systems can be stable, and they can be predictable, but they can also change and display unintended or unexpected behaviours. Thinking in systems requires understanding that ‘everything is connected to everything else’ and that ‘you can’t just do one thing’.⁵³ The boundaries of a system and the emergent properties a system exhibits need to be determined backwards from observation of it in action, rather than forwards from its specification and implementation.

1.2. Machines that Learn

Over the course of researching for and writing this thesis, machine learning has moved from a scientific term of art to a minor household name. Once restricted to a somewhat marginalised research community, applied machine learning is now experienced by virtually all regular internet users through content personalisation on the basis of externally-held databases and information leaked by browsers, apps, hardware and software protocols. Indeed, while recognition of the term machine learning might seem low (in 2016) at one-in-ten people in the UK,⁵⁴ in relation to particular applications, such as voice recognition or web personalisation, awareness among young people at least is significantly higher, ranging from 48%–85% depending on application area according to a survey of 2,044 individuals undertaken for the Wellcome

⁵² Nick Seaver, ‘Knowing Algorithms’ in *Paper presented at Media in Transition 8, Cambridge, MA* (2013) 10.

⁵³ John D Stermann, *Business Dynamics* (McGraw-Hill 2000) 4.

⁵⁴ Ipsos MORI, *Public views of Machine Learning* (The Royal Society 2017).

Trust.⁵⁵

Machine learning encompasses a set of techniques to *discern and operationalise patterns in data*. Machine learning approaches are about building models without explicitly programming them, by inducing them from specific datapoints. A variety of different machine learning techniques use different approaches to extract and encode patterns that are likely to generalise well to examples not present in the data used to train the system.

Different flavours of machine learning exist. *Supervised* learning refers to a machine learning approach to building a model to predict value(s) or classification(s) from input data, given a training dataset where both input data and ‘correct’ value(s) or classification(s) (known as ‘ground truth’) are present. The ‘supervised’ aspect does not refer to human oversight, but the way that the labels in the historical data guide the machine learning approach to map inputs to outputs in a way consistent with the training set. A canonical example of this type of learning would be to build a model to determine whether unseen individuals are likely to default on their mortgage within the next six months, based on observed customer data.

Unsupervised learning sits in contrast to supervised approaches, and is used to extract patterns about data lacking clear labels or outcomes. For example, data on trips taken using public transport cards could be clustered to see what types of typology of users such data supports. Such approaches in fields such as image recognition might not be able to tell users which picture contains a dog and which a cat, but it might be able to discern that there are two main groups of animals in photographs. A famous example of a deployed system which can be considered as unsupervised learning is Google’s original search algorithm, *PageRank*, which evaluated webpages based on the quality and quantity of resources which linked to it, not requiring preset labels or outcomes in historical data.

Other flavours exist and are worth highlighting, although these do not take centre stage in this thesis. *Semi-supervised* learning is appropriate when some labels are present, but instead of discarding data lacking labels, aspects of its structure can be used in addition to improve the model. *Reinforcement* learning approaches involve a machine agent taking actions in an environment that are associated with rewards, and seeking to maximise these rewards, particularly over the long or medium term. This differs from supervised learning, as the data need not be presented in advance, and agents must also trade off by balancing *exploring* their environment with *exploiting* it to gain the most cumulative reward. These approaches have not been heavily deployed

⁵⁵ Rebecca Hamlyn, Peter Matthews and Martin Shanahan, *Science Education Tracker: Young people’s awareness and attitudes towards machine learning* (The Wellcome Trust, the Royal Society, and the Department for Business, Energy & Industrial Strategy 2017) DOI: 10/cxnf.

1. Hello, World!

in practice, but have received considerable attention due to the media-friendly advances in research seen as reinforcement learning systems successfully learn to play video games in simulated environments.

Even within these technologies, many different families of machine learning approaches exist. Some of them date back decades—or arguably, even over a century. In my experience, it has often surprised researchers outside the field to learn that linear regression, a common tool in many disciplines, is itself a form of machine learning: presumably because they thought this new technology must be much further from their own skills, experience and understanding than it often is.⁵⁶

Despite being a technique effectively discovered in the late 1800s by Francis Galton,⁵⁷ linear modelling is commonly taught in the introductory sessions of university machine learning courses. The most well-known form of linear model, linear regression, is an example of a statistical model which, by making some heavy assumptions about the distributions and structure of the patterns in a dataset, attempts to represent the processes by which the data was generated. In multivariate regression, a hyperplane is fitted to data using the method of *least squares*, whereby the coefficients defining the hyperplane are chosen as to minimise the residual sum of squares between the plane and the datapoints. Add a new group of datapoints, optimise the hyperplane to the residual sum of squares again, and the machine has ‘learned’ a different model. While linear regression models built this way are often put to *explanatory* uses, through using models’ coefficients to test hypotheses relating to correlations between the dependent and independent variables in the model, they can also be used for prediction. Give me a vector omitting only one component, and I can use the hyperplane to predict its value using all the others.

Other machine learning approaches exist. Some resemble augmented versions of regression (eg *lasso regression*⁵⁸) or linear models in implicit high dimensional feature spaces (eg *support vector machines with the kernel trick*⁵⁹). Others rely on building one or more decision trees using the features in a dataset (eg *CART, random forests*).⁶⁰ Perhaps the only approach to machine learning with some anchoring in popular culture

⁵⁶ Memorably, a highly decorated and very amicable natural scientist I worked with on a policy report on this very topic, upon being told this in a meeting and, after looking to the tenured statistician opposite for confirmation (which she received with a nod) exclaimed with incredulous surprise: ‘Regression? Even I can do that!’.

⁵⁷ Jeffrey M Stanton, ‘Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors’ (2001) 9(3) *Journal of Statistics Education* DOI: 10/gd82dx.

⁵⁸ Robert Tibshirani, ‘Regression Shrinkage and Selection via the Lasso’ (1996) 58(1) *Journal of the Royal Statistical Society. Series B (Methodological)* 267.

⁵⁹ Nello Cristianini and John Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge University Press 2000).

⁶⁰ Leo Breiman, ‘Random Forests’ (2001) 45(1) *Machine Learning* 5 DOI: 10/d8zjwq.

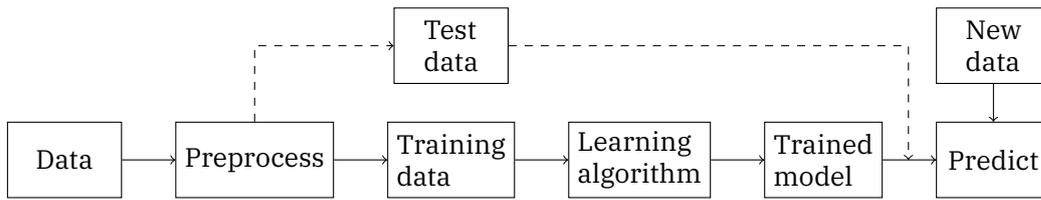


Figure 1.1.: The machine learning pipeline. Diagram by author.

and consciousness is the *neural network*.⁶¹ The neural network approach has a long history that dates back to theorising in the 1940s and the first successful computers running learning systems such as simple ‘perceptrons’ in the 1950s.⁶² At the end of the 1960s, the book *Perceptrons* by researchers Marvin Minsky and Seymour Papert notoriously claimed that not only did perceptrons suffer a critical computational flaw that left them unable to compute certain aspects of predicate logic, such as XOR, but that these flaws would apply to all neural networks, amending them would leave them computationally nigh impossible to calculate.⁶³ Since then, findings (notably the discovery of *backpropagation*) that have nullified these critiques have been discovered.

In recent years, neural networks with a great number of successive layers have shown high skill in some areas, such as certain image recognition tasks, which were previously thought to be computationally very challenging. These types of neural networks systems have become known collectively known as *deep learning* approaches, and have been involved in many of the more flashy and surprising successes in the machine learning field over the last decade.

1.3. Making the Machines that Learn

Despite media framings of automated, thinking systems, training a machine learning model is much more involved than pointing a piece of software at a dataset and pressing a large, red button marked ‘learn!’. The role of humans in designing and maintaining these systems is usually heavily underplayed to accentuate ‘the AI’s’ achievement.

For example, IBM’s ‘Watson’ system, famous for beating the incumbent champions on the American gameshow *Jeopardy*, brands itself in an anthropomorphic way, akin to an ‘artificial general intelligence’⁶⁴ algorithm like *HAL 3000* from *2001: A Space Odis-*

⁶¹ The character *Data* in *Star Trek: The Next Generation* (1987–94), for example, is often referred to as being powered by a ‘neural net’.

⁶² Neha Yadav, Anupam Yadav and Manoj Kumar, *An Introduction to Neural Network Methods for Differential Equations* (Springer 2015) DOI: 10/gfgrb3 13–14.

⁶³ Marvin Minsky and Seymour Papert, *Perceptrons* (MIT Press 1969).

⁶⁴ It should be noted that work is not concerned with artificial general intelligence (AGI). Many of the ques-

1. Hello, World!

sey. Yet the reality is considerably less autonomous: different mathematical systems are strung together with both other software systems and with humans⁶⁵ for in different ways for different tasks. If there is a ‘general’ part of these systems’ ‘general intelligence’, it is broadly a result of the flexible human resources that build them. Google’s laboratory subsidiary DeepMind, responsible for much recent cutting edge research and investment in the machine learning space, has made their internal machine learning toolkit *TensorFlow* open-source. Alongside the extreme brain-drain drawing academics into the private sector⁶⁶ and the generally open publishing record of the latest research into machine learning systems at global conferences,⁶⁷ this emphasises the relative importance of people (and data) compared to learning algorithms themselves.

Furthermore, humans are key even within narrowly defined machine learning projects. This becomes more apparent by considering a more fleshed-out definition of machine learning from American computer scientist Tom Mitchell:

A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks *T*, as measured by *P*, improves with experience *E*.⁶⁸

Here, the designer is in the driving seat. They alone define what constitutes a valid *task*, valid *experience*, and relevant *performance measure*.

Tasks that might seem well-defined during conversation will often need further formalisation before they are suited to computerisation. A seemingly straightforward task to ‘predict the firms most likely to become insolvent’ needs considerable further specification. Over what timeframe is the prediction of interest—a week, a month, a year? How is insolvency defined and measured—when it is reported publicly, or when it occurs privately? Is this modeling over all firms of all sizes, or over firms of particular value that are specifically sectorally or spatially situated? These issues are heavily

tions within are equally as applicable to high dimensional regression methods as to neural networks. For work and some concerns surrounding AGI, see eg Nick Bostrom, *Superintelligence: Paths, dangers, strategies* (Oxford University Press 2014).

⁶⁵ For example, crowdworkers are commonly employed within the ‘automated’ workflows, particularly for difficult tasks. See Olivia Solon, ‘The Rise of ‘Pseudo-AI’: How Tech Firms Quietly Use Humans to Do Bots’ Work’ (*The Guardian*, 6th July 2018) (<https://perma.cc/Q4P9-ZCHF>). For a tongue in cheek perspective, see Randall Munroe, ‘Self Driving’ (*xkcd webcomic*, 2017) (<https://xkcd.com/1897/>).

⁶⁶ Amir Mizroch, ‘Artificial-intelligence experts are in high demand’ (*The Wall Street Journal*, 1st May 2015); Elizabeth Gibney, ‘AI talent grab sparks excitement and concern’ (2016) 532(7600) *Nature* 422 DOI: 10/bfrc.

⁶⁷ See eg Doug Eck, ‘NIPS 2016 & Research at Google’ (*Google Research Blog*, 4th December 2016) (<https://research.googleblog.com/2016/12/nips-2016-research-at-google.html>).

⁶⁸ Tom M Mitchell, *Machine learning* (McGraw Hill 1997). Credit also goes here to Mireille Hildebrandt, who demonstrated how this definition is an important tool in fostering discussions on machine learning across disciplinary boundaries. See eg Mireille Hildebrandt, *Smart technologies and the End(s) of Law* (Edward Elgar 2015).

subjective and value-laden,⁶⁹ and considering them as mere technicalities shuts down these important political discussions.

Experience—the input data for a model—is similarly subjective. Where data is collected, there are decisions to be made about who, and how, to sample from the world. While ‘big data’ proponents have argued that we are entering a time where sampling is unnecessary, as $n = all$,⁷⁰ critics have pushed back firmly against this, noting that not only does this ignore the fact that ‘quantification always implies a preceding qualification’, but it does not reflect the necessary incompleteness of knowing the world in a quantified way: ‘N is not All and All is not N’.⁷¹

Qualification happens not just in the act of measuring and sampling, but when pre-processing the datasets collected. Data from real world sources almost always requires cleaning: processes consisting of correcting, removing or imputing missing or non-standard answers; grouping similarly named inputs; changing the format and shape of the data storage; dealing with differences in format and capitalisation; detecting and correcting erroneous data entries, and so on. Around 80% of the time that data analysis takes is thought to be spent on these kinds of issues.⁷² The decision on what is ‘clean’ or ‘unclean’ is a human one, and as data can be cleaned or classified in more than one way, this a highly value-laden task often obscuring important moral questions.⁷³ Classifying cases that do not fit clearly into taxonomies is clearly a heavily value-laden task. Similarly, there is no clear answer around what to do when data is missing. Machine learning systems often require a complete table of data in order to run the model. Discarding incomplete records is possible—as long as those records were *missing-at-random*. If there is a correlation between missing data and some phenomenon of interest, then there is a severe risk of biasing analysis based on that dataset. Other methods, such as *imputing* a best-guess of what that data would have been, require similarly considerable assumptions.⁷⁴

Machine learning systems often additionally require a design stage known as ‘fea-

⁶⁹ See Peter Miller and Ted O’Leary, ‘Accounting and the construction of the governable person’ (1987) 12(3) *Accounting, Organizations and Society* 235 DOI: 10/d3cg8w, on the concept of creating a ‘governable person’ through practices such as credit scoring; Martha Poon, ‘From new deal institutions to capital markets: Commercial consumer risk scores and the making of subprime mortgage finance’ (2009) 34(5) *Accounting, Organizations and Society* 654 DOI: 10/cm8g3x, on credit scoring interpretations were changed not for the purposes of accuracy but for the purposes of action, enabling the ‘subprime’ crisis.

⁷⁰ Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work and Think* (Hodder & Stoughton 2013).

⁷¹ Mireille Hildebrandt, ‘Slaves to Big Data. Or Are We?’ (2013) 17 *IDP Revista de Internet Derecho y Política* 27 DOI: 10/gd82jr.

⁷² Tamraparni Dasu and Theodore Johnson, *Exploratory data mining and data cleaning* (John Wiley & Sons 2003); Hadley Wickham, ‘Tidy Data’ (2014) 59(1) *Journal of Statistical Software* 1 DOI: 10/gdm3p7.

⁷³ Bowker and Star (n 35); boyd and Crawford (n 19).

⁷⁴ Andrew Gelman and Jennifer Hill, ‘Missing-Data Imputation’ in *Data Analysis Using Regression and Multi-level/Hierarchical Models* (Cambridge University Press 2006) DOI: 10/dv8cb9.

1. Hello, World!

ture engineering’, where available input variables are combined or transformed into the data that eventually feeds the model. This process surrounds manipulating the input data into variables that are more likely to be predictively useful or computationally efficient. For example, preprocessing is required to transform data containing a work postcode and a home postcode into a distance between the two. Without feature engineering, machine learning algorithms would almost certainly lack the skill to create useful information from an untransformed postcode. If treated solely as a series of three characters and three numbers, as a naïve algorithm might, a UK postcode alone has over five hundred million possible permutations, with the numbers representing historical artefacts without geographic logic.⁷⁵

Feature engineering depends strongly on the type of data being fed into a system. When pictures are used, transformations are undertaken including how dark or light an image is, characteristics of an image when flipped, and so on. Text data are often transformed in advance into a high dimensional vector using specialist unsupervised learning systems.⁷⁶ At this stage of the design process, the humans involved are choosing some features of interest, while omitting to choose others. They are usually doing so on a basis of ‘what works’ and ‘what is my task’—combining heuristic ‘folk-knowledge’ of machine learning⁷⁷ with experimentation in application specific contexts. *Deep learning* approaches differ mostly from existing forms of machine learning insofar as they seek to automate (and by extension optimise) feature engineering. While this does not always work (for example, where, as with postcodes, external context matters), for many forms of data, particularly those where each variable has a relatively abstract meaning such as pixels or audio signal, this has produced significant performance increases.

Measuring this performance however presents similarly value-laden choices. This is unsurprising when considering that machine learning seeks to choose the ‘best’ model: and humans have often struggled to conclude what is better and worse with any degree of consensus or stability. *Loss functions*, which quantify the degree of disagreement between a prediction and actual outcome, are at the core of many issues of per-

⁷⁵ For example, the number on the end of the first part of a London postcode is number of the historical district post office in alphabetical order, apart from the first which represents the head post office for that part of the city.

⁷⁶ Word embedding systems such as *word2vec* Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean, ‘Distributed Representations of Words and Phrases and their Compositionality’ in CJC Burges, L Bottou, M Welling, Z Ghahramani and KQ Weinberger (eds), *Advances in Neural Information Processing Systems 26* (2013) and *GloVe* Jeffrey Pennington, Richard Socher and Christopher D Manning, ‘GloVe: Global vectors for word representation’ in *Empirical Methods in Natural Language Processing (EMNLP)* (2014) (<http://www.aclweb.org/anthology/D14-1162>) are among the most popular tools used for this purpose.

⁷⁷ Pedro Domingos, ‘A few useful things to know about machine learning’ (2012) 55(10) *Commun. ACM* 78 DOI: 10/cgc9.

formance within a machine learning model. ‘Best’ is defined as minimising this loss function. Some in classification are simple—one point added for every error made, for example. With continuous data, it becomes necessary to set a scoring system. When is a close guess considered an error? Is the algorithm penalised more, the further away its guess is? In that case, one wildly incorrect prediction could disqualify an algorithm that performs very well on all other observations. Are different misclassifications penalised in different ways? Perhaps thinking A was actually B would not cause much of a problem in practice, but mistaking B for A would be catastrophic. Automated car brakes should likely err on the side of caution in the face of an uncertain collision, but not so much that precautious braking becomes a hazard in and of itself.

In sum, algorithmic systems and machine learning in particular involve humans and values *even* with only a cursory glance at the processes involved in making and managing them on paper. I now turn to how their value-ladenness has caught the attention of academics, journalists and the general public.

1.4. Algorithmic War-Stories

Algorithmic systems involving machine learning have been subject to a considerable amount of recent furore. The research field concerning the social aspects and implications of machine learning field has, perhaps more than many others, been motivated by stories reported in the media. Journalists in turn have been motivated by a call for arms suggesting they can and should be a check and balance whilst regulatory regimes develop and mature.⁷⁸ The ‘algorithmic war-stories’⁷⁹ that have resulted from this are important scene-setters. Even if some of them sit amongst the ‘academic urban legends’ that get uncritically reproduced in much modern scholarship,⁸⁰ or propagate unjustified ‘algorithmic drama’,⁸¹ they do reflect and contextualise the types of alleged social harm that have engaged researchers, and serve to contextualise both these debates and the thesis as a whole, particularly for those reading it in the future where issues focussed upon may significantly differ.

⁷⁸ Nicholas Diakopoulos, ‘Algorithmic accountability’ (2014) 3(3) *Digital Journalism* 398 DOI: 10/gc5t4g. See also the new journalistic initiatives focussing on algorithmic systems such as The Markup (<https://themarkup.org/>) in the US and Algorithm Watch (<https://algorithmwatch.org>) in Germany.

⁷⁹ This term was coined by my friend and collaborator Lillian Edwards, and used extensively in Lillian 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 L. & Tech. Rev.* 18 DOI: 10/gdxthj.

⁸⁰ Ole Bjørn Rekdal, ‘Academic Urban Legends’ (2014) 44(4) *Social Studies of Science* 638 DOI: 10/gd89bc.

⁸¹ Malte Ziewitz, ‘Governing Algorithms: Myth, Mess, and Methods’ (2016) 41(1) *Science, Technology, & Human Values* 3 DOI: 10/gddv9k.

1.4.1. Recidivism and Racism

A significant motivator for the academic interest in machine learning was the 2016 report by American investigative journalism outfit ProPublica into recidivism prediction instruments (RPI) deployed in varying parts of the US justice system.⁸² Developed by Northpointe Inc (now equivalent), the Correctional Offender Management Profiles for Alternative Sanctions (COMPAS) system assigns risk scores for varying types of recidivism. These scores are used by judges, probation and parole officers, among others, to inform their decisions. ProPublica looked at Broward County in Florida, a large jurisdiction primarily using the COMPAS tool to determine whether to release or detain a defendant before his or her trial. Through a public records request, the journalists received two years worth of COMPAS scores from the Broward County Sheriff's Office, covering 18,610 individuals scored in 2013 and 2014, 11,757 of whom were assessed at the pretrial stage. Using first names, last names and dates of birth, they matched individuals to criminal records available from the Broward County Clerk's Office website, and aligned the scores received with the offences they believed they related to. They sought to determine if a person had been charged with a new crime subsequent to that for which they were COMPAS scored, establishing a ground truth for recidivism.⁸³

The journalists found that the likelihood of a black defendant who would *not* go on to recidivate being classified as high risk is almost twice that of similar white defendants. Additionally, the likelihood of a recidivating black defendant receiving a low risk assessment is around half that of similar white defendants. In statistical terms, this translates to the finding that black defendants have higher *false positive rates* and lower *false negative rates* than their white counterparts. 'There's software used across the country to predict future criminals', led the ProPublica report. 'And it's biased against blacks.'

A further issue with these systems resulted from their proprietary nature. While ProPublica could, through laws allowing access to public records, obtain enough data to analyse the system, they could not access the system itself in order to scrutinise it fully.⁸⁴

Statistically, the analysis of the situation soon proved a little more thorny. At the time of publication, the lead reporter on this story did caution on Twitter that 'the statistical

⁸² Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, 'Machine bias' (*ProPublica*, 23rd May 2016) (<http://perma.cc/L4M4-TJQT>).

⁸³ Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, 'How We Analyzed the COMPAS Recidivism Algorithm' (*ProPublica*, 23rd May 2016) (<https://perma.cc/W3EB-BKW4>).

⁸⁴ Interestingly, in the UK it is highly likely that they would not have been able to access this data due to data protection being an important grounds for refusing to fulfil a freedom of information request.

analysis for [the] story on criminal prediction algorithms is insanely wonky'.⁸⁵ This was soon a topic of academic debate, as researchers showed that under certain conditions, the measure of fairness used by Northpointe Inc was not statistically compatible with the measure of fairness used by ProPublica.⁸⁶ ProPublica journalists were concerned about *disparate mistreatment*, where error types (eg false positives or false negatives) disproportionately affect a salient or protected population subgroup.⁸⁷ This seems intuitively undesirable. As well as avoiding disparate mistreatment, it seems desirable that a risk score is *calibrated* correctly. This means that if we consider a group of individuals that the model says have a certain probability of recidivating—say, a 70% chance—then we would like about 70% of that group to actually turn out to recidivate. Furthermore, if there are different population subsets, such as ethnic groups, we would like this calibration condition to be maintained simultaneously for each of these sets of individuals. If calibration across groups is *not* maintained in a risk scoring system, then the probabilities given by the machine would demand different interpretations depending on the group, which risks incentivising users to take protected characteristics into account during the interpretation of the predictions.⁸⁸ However, calibration and the avoidance of disparate mistreatment are at tension in cases where the prevalence to recidivate does truly differ between ethnic groups taken in aggregate.⁸⁹ While ProPublica did not perhaps uncover the smoking gun they thought they did, as often is the case with complex issues such as these, they stumbled upon an even more interesting quandary.

⁸⁵ Julia Angwin (*Twitter* [*@JuliaAngwin*], 23rd May 2016) (<https://perma.cc/6QLB-KRFM>) accessed 27th November 2017.

⁸⁶ Alexandra Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments' (2017) 5(2) *Big Data* 153 DOI: 10/gdcdqf; Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores' in Christos H Papadimitriou (ed), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik 2017) vol 67 DOI: 10/gfgq8s.

⁸⁷ Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez and Krishna P Gummadi, 'Fairness beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment' in *Proceedings of the 26th International Conference on World Wide Web* (International World Wide Web Conferences Steering Committee 2017) DOI: 10/gfgq8r.

⁸⁸ Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg and Kilian Q Weinberger, 'On Fairness and Calibration' in I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett (eds), *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc 2017).

⁸⁹ Chouldechova (n 86); Kleinberg, Mullainathan and Raghavan (n 86). It is worth noting here that this casts the spotlight heavily on how such groups are constructed, pointing to the importance of sociology and philosophy of science research on classification such as Bowker and Star (n 35); Hacking, 'Kinds of People' (n 41).

1.4.2. Discrimination and Search Engines

In 2013, Latanya Sweeney, a security researcher at Harvard University, did something many academics are prone to do. She Googled her own name. Instead of her latest citations or media quotes (she was looking for a paper to show a reporter⁹⁰), something else stood out to her. Besides the results, the ‘sponsored’ adverts tailored to her search asked a simple question: ‘Latanya Sweeney, Arrested?’.

She had not been arrested, but, exploring further, she noted that other adverts of this type would appear with minor search modifications, offering services such as legal help, or the ability to search someone’s criminal history. As a researcher specialising in investigating real world systems (she made her name in part through a study showing that public and semi-public hospital discharge data could be easily re-identified for a large proportion of the population using voter records that could be purchased for \$20⁹¹), Sweeney sought to investigate this. By establishing a list of predominantly ‘white-sounding’ names (eg Allison, Brendan, Brett) and ‘Black-sounding’ names (eg Darnell, Keisha, Jermaine), drawing on prior research in this area, and testing 2,184 of them in Google Search, she found that the latter generated a far higher percentage of adverts associated with or using the word *arrest* when compared to ads delivered to ‘white-sounding’ first names. On one of the two websites examined, a black-identifying name was 25% more likely to get an ad suggestive of an arrest record. Sweeney also ruled out knowledge of any criminal record of the person to whom the ad was delivered.⁹² Acknowledging that it was beyond the scope of her research to know what was happening in the ‘inner workings of Google AdSense,’ and whether the apparent bias displayed was the fault of society, Google or the advertiser, Sweeney still asserted her research raised questions about society’s relationship to racism and the role of online advertising services in this context.

In an even earlier incident of notoriety in 2004, the Google search algorithm(s) placed a site *Jew Watch* at the top of the rankings for many people who searched for the word ‘Jew.’ Google (in stark contrast to its more recent attitudes⁹³) refused to manually alter their ratings and claimed instead that the preferences of a particular group of searchers had put Jew Watch to the top rather than any normative ranking by Google. It was stated that ‘[B]ecause the word ‘Jew’ is often used in an anti-Semitic context,

⁹⁰ Latanya Sweeney, ‘Saving Humanity’ in *1st Conference on Fairness, Accountability and Transparency (FAT*)*, New York, 23 February (Keynote address) (2018) (https://youtu.be/OIK_nVOM2tc).

⁹¹ Latanya Sweeney, ‘Simple Demographics Often Identify People Uniquely’ [2000] Carnegie Mellon Univ. Data Privacy Working Paper 3 (<https://perma.cc/6V9N-9QNG>).

⁹² Latanya Sweeney, ‘Discrimination in Online Ad Delivery’ (2013) 11(3) Queue 10 DOI: 10/gdxwj6.

⁹³ Google has rethought its approach to such cases, especially after unfavourable press reports. Now, a ‘quality rating’ downgrades pages rather than removes them. See Samuel Gibbs, ‘Google Alters Search Autocomplete to Remove ‘are Jews Evil’ Suggestion’ (*The Guardian*, 5th December 2016) (<https://perma.cc/P4GF-HTQB>).

this had caused Google's automated ranking system to rank *Jew Watch*—apparently an anti-Semitic web site—number one for the query.⁹⁴ In the end Google refused to remove the site from the rankings but collective effort was encouraged among users to push up the rankings of other non-offensive sites.

1.4.3. Know Your Customers

The ways in which firms profile individuals, such as customers or job adverts, has long been of concern to scholars, and more recently to the public.

One well-known war-story concerns the profiling practices of the American supermarket Target. A magazine piece, now urban legend, claimed a teenage daughter was targeted with pregnancy related offers, based on loyalty card data, through the mail. These offers were seen by the father of the family, triggering an awkward conversation before the expectant mother had been ready to have it.⁹⁵

The details of this story are unclear, and whether it is true at all has been questioned.⁹⁶ Yet regardless of the veracity of this particular incident, loyalty card data has long been used for tracking and profiling. Two decades ago already, data from UK supermarket Tesco's 'Clubcard' scheme formed the basis for 80,000 variants of a single shot of direct marketing material.⁹⁷ In other instances, retailers have even sought to link credit card details back to customers' addresses by employing 'reverse append' practices in order to build profiles, bringing a slew of specific legal and legislative response.⁹⁸

In-store tracking is significantly and increasingly concerning to a large proportion of consumers. According to a 2014 survey of 1,042 US consumers conducted by American consumer feedback company OpinionLab, 80% of respondents find in-store tracking using mobile phones unacceptable, and 81% said they don't trust retailers to keep data private and secure.⁹⁹ A study for the European Commission in 2015 reports that tracking concerns in retail contexts are growing in the UK, with 45% (of 1,328) of UK residents concerned about tracking via loyalty cards (the third most concerned nation in Europe, up from 36% in 2005) and 51% concerned about being recorded in

⁹⁴ Danny Sullivan, 'Google In Controversy Over Top-Ranking For Anti-Jewish Site' (*Search Engine Watch*, 24th April 2004) (<https://perma.cc/8ZCC-7WFB>).

⁹⁵ Charles Duhigg, 'How Companies Learn Your Secrets' (*New York Times Magazine*, 16th February 2012) (<https://perma.cc/2E69-JRKW>).

⁹⁶ See for example scepticism over the story's veracity here <https://perma.cc/FE9Y-CT6J>.

⁹⁷ Margon Georgiadis, Katrina Lane and Sue Whalley, *Smart data, smart decisions, smart profits: The retailers' advantage* (McKinsey & Company 2000) (<https://perma.cc/7T5W-WSYB>).

⁹⁸ Glenn A Blackmon, 'Problems at the register: Retail collection of personal information and the data breach' (2014) 65 Case W. Res. L. Rev. 861.

⁹⁹ OpinionLab, New study: consumers overwhelmingly reject in-store tracking by retailers (March 2014) (<https://perma.cc/R289-HZ3Z>).

1. Hello, World!

private, commercial spaces (the most concerned nation in Europe, up from 40% in 2005).¹⁰⁰

In online commerce too, profiling has received criticism effectively since its emergence. More broadly, cases around ‘redlining’ on the internet—‘weblining,’ as it was known nearly 20 years ago—are far from new.¹⁰¹ A spate of stories in 2000 during the heady years of the dot-com bubble surrounded racist profiling using personal data on the internet. Consumer bank *Wells Fargo* had a lawsuit filed against it for using an on-line home-search system to steer individuals away from particular districts based on provided racial classifications.¹⁰² Similarly, the online 1-hour-media-delivery service *Kozmo* received a lawsuit for denying delivery to residents in black neighbourhoods in Washington, DC, which they defended in the media by saying that they were not targeting neighbourhoods based on race, but based on high Internet usage.¹⁰³

In more modern tracking and personalisation systems online, targeted adverts have also been accused of having discriminatory effects. Two examples are illustrative of this. Firstly, researchers using an automated experimentation system discovered that where users identified as male, they received more adverts encouraging the seeking of coaching services for high paying jobs more than those identifying as female, despite identical browsing patterns representing a job search.¹⁰⁴ This exacerbates concerns around how intermediating platforms may exacerbate the gender pay gap—and the important role of these actors is not anticipated for well in existing US law.¹⁰⁵ A similar, parallel concern emerged around how adverts can be targeted on social media network *Facebook*. In particular, while direct discrimination on the basis of race in context such as housing adverts is forbidden under US law, *Facebook* was revealed by journalists at *ProPublica* to allow targeting on the basis of so-called ‘ethnic affinity’,¹⁰⁶ and continued to do so for more than a year despite this practice having been clearly flagged and well-publicised.¹⁰⁷

¹⁰⁰ European Commission, Special Eurobarometer 431: “Data Protection” (European Union 2015) DOI: 10.2838/552336.

¹⁰¹ Marcia Stepanek, ‘Weblining: Companies are using your personal data to limit your choices—and force you to pay more for products’ [2000] *Bloomberg Business Week*, 26.

¹⁰² Anonymous, ‘Wells Fargo yanks “Community Calculator” service after ACORN lawsuit’ [2000] *Credit Union Times* (<https://perma.cc/XG79-9P74>).

¹⁰³ Elliot Zaret and Brock N Meeks, ‘Kozmo’s digital dividing lines’ [2000] *MSNBC*; Kate Marquess, ‘Redline may be going online’ (2000) 86 *ABA J.*, 81.

¹⁰⁴ Amit Datta, Michael Carl Tschantz and Anupam Datta, ‘Automated Experiments on Ad Privacy Settings’ (2015) 2015(1) *Proceedings on Privacy Enhancing Technologies* 92 DOI: 10/gcv7m7.

¹⁰⁵ Amit Datta, Anupam Datta, Jael Makagon, Deirdre K Mulligan and Michael Carl Tschantz, ‘Discrimination in Online Advertising: A Multidisciplinary Inquiry’ in *Conference on Fairness, Accountability and Transparency* (2018) (<http://proceedings.mlr.press/v81/datta18a.html>).

¹⁰⁶ Julia Angwin and Terry Parris Jr, ‘Facebook Lets Advertisers Exclude Users by Race’ (*ProPublica*, 28th October 2016) (<https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race>).

¹⁰⁷ Julia Angwin, Ariana Tobin and Madeleine Varner, ‘Facebook (Still) Letting Housing Advertisers Ex-

1.4.4. Bias in the Toolkit

In addition to the narrative around machine learning system making decisions, it is important to see machine learning's emerging role as *components* of broader software systems: as a tool for a specific task. Within these contexts, two tools which are often intermediate steps in designing a broader application have revealed public concern: *image recognition* and *natural language processing*.

Probably the most public failure of image recognition technology was spurred by a tweet in June 2015 by software developer Jacky Alciné.¹⁰⁸ This tweet included a photo where Mr Alciné and his friend, both Black, were identified as 'gorillas' by *Google Photos*. The firm responded quickly on *Twitter*, as one of their engineers described it as 'high on [their] list of bugs you *never* want to see happen',¹⁰⁹ and apologised publicly.¹¹⁰ Even years later, the fix that *Google* concluded appropriate was not to deal fully with any underlying prediction errors in the system, but to remove the model's ability to predict gorillas *entirely*.¹¹¹

While the Google system sought to identify objects and misclassified faces, some systems are designed to recognise faces in the first place, whether to infer characteristics from them (such as emotions or demographics) or to ascertain or verify an individual's identity. These have attracted a related strain of criticism—that commercial facial analysis systems disproportionately fail on people of colour, in particular, women of colour.¹¹² This in turn has attracted criticism on social media and at workshops, focussing largely on the point that oversurveilled demographic groups should not seek to be surveilled with equal accuracy: and so 'fixing' these systems brings problems in and of itself.¹¹³

Natural language processing has similarly challenging problems. One of the most common tools used today is the *word embedding*. Word embedding is an approach where words are mapped to vectors of real numbers, and their geometric position in relation to one another is used to understand something about their semantic meaning. They have a significant history, particularly from the 1970s onwards, where vector space models were used to map documents. Similar documents would be located

clude... — ProPublica' (*ProPublica*, 21st November 2017) (<https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin>) accessed 1st October 2018.

¹⁰⁸ Jacky Alciné (*Twitter* [*@jackyalcine*], 28th June 2015) (<https://perma.cc/E6UT-K8GL>).

¹⁰⁹ Yonatan Zunger (*Twitter* [*@yonatanzungger*], 28th June 2015) (<https://perma.cc/7PMP-ZLT9>).

¹¹⁰ Jana Kasperkevic, 'Google Says Sorry for Racist Auto-Tag in Photo App' (*The Guardian*, 1st July 2015) (<https://perma.cc/A24K-ZXV6>).

¹¹¹ Tom Simonite, 'When It Comes to Gorillas, Google Photos Remains Blind' (*WIRED*, 11th January 2018) (<https://perma.cc/L52Z-J8J6>).

¹¹² Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' in *Conference on Fairness, Accountability and Transparency* (2018) (<http://proceedings.mlr.press/v81/buolamwini18a.html>) accessed 1st October 2018.

¹¹³ See eg the question and answer session for *ibid* at <https://www.youtube.com/watch?v=Af2VmR-iGkY>.

1. Hello, World!

close to each other, and distinct documents located further away.¹¹⁴ The approach took off significantly in the early 2010s, where researchers created new and more efficient means to create such vectors using large corpora of text such as web crawls and neural network methods, resulting in tools such as *word2vec* from a team at Google¹¹⁵ and *GloVe* from researchers at Stanford.¹¹⁶ While the research includes the tools used to make word vectors, because of the difficulty and computational intensity in training them, most users use the output of these models as a tool, which effectively takes the form of pre-trained large tables of data. In these tables, each row represents a word and each column (of which there can be hundreds) records a number. Each word can then be seen as sitting as a point in the n -dimensional space, where n is the number of columns in the word embedding matrix.

The geometric relations between these words in the n -dimensional space exhibit interesting properties. For example, the vector between the words ‘man’ and ‘woman’, when it is projected onto the word ‘king’, approximates the word ‘queen’.¹¹⁷ Unfortunately, language contains unsavoury relations too, particularly when it is picked up from varied, societally mediated and messy data, as highlighted primarily in two studies from 2016–17. One study showed the existence of a gendered dimension in word embeddings replicating stereotypes: that not only is *man* to *woman* as *king* is to *queen*, but also as *architect* is to *interior designer*, *football* is to *volleyball*, or *computer programmer* is to *homemaker*.¹¹⁸ A separate concurrent study created a statistical analogue to the Implicit Association Test used to understand word relations in humans, highlighting recoverable imprints of historic human biases as also present in word embeddings.¹¹⁹ These representational issues, or semblances of them, appear to be translated into common tools, such as Google Translate.¹²⁰ Given word embeddings are

¹¹⁴ G Salton, A Wong and CS Yang, ‘A Vector Space Model for Automatic Indexing’ (1975) 18(11) Commun. ACM 613 DOI: 10/fw8vv8; PD Turney and P Pantel, ‘From Frequency to Meaning: Vector Space Models of Semantics’ (2010) 37 Journal of Artificial Intelligence Research 141 DOI: 10/gd85zk.

¹¹⁵ Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, ‘Efficient Estimation of Word Representations in Vector Space’ [2013] arXiv preprint (<https://arxiv.org/abs/1301.3781>).

¹¹⁶ Jeffrey Pennington, Richard Socher and Christopher Manning, ‘GloVe: Global Vectors for Word Representation’ in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014).

¹¹⁷ Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig, ‘Linguistic Regularities in Continuous Space Word Representations’ in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2013).

¹¹⁸ Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama and Adam Kalai, ‘Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings’ [2016] 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain; Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama and Adam Tauman Kalai, ‘Quantifying and Reducing Stereotypes in Word Embeddings’ [2016] arXiv preprint, on methods to remove this latent gender dimension.

¹¹⁹ Aylin Caliskan, Joanna J Bryson and Arvind Narayanan, ‘Semantics derived automatically from language corpora contain human-like biases’ (2017) 356(6334) *Science* 183 DOI: 10/f93cpf.

¹²⁰ Brian Resnick, ‘How Artificial Intelligence Learns to Be Racist’ [2017] *Vox* (<https://perma.cc/4ZJP-HB3G>) accessed 2nd October 2018.

heavily used in many software systems, the biases uncovered by these researchers have been thought to risk issues downstream in areas such as CV analysis for job application filtering.

1.4.5. Was It Facebook Wot Won It?

During the course of researching for and writing this thesis, one news piece stood out more than any other. In 2017–18, an array of investigative news stories largely from The Observer shone light on the microtargeting practices of a range of firms on behalf of international political campaigns.¹²¹ The most notorious firm implicated was Cambridge Analytica, accused of ‘weaponising’ a large array of data on individuals and their Facebook friends collected by a personality quiz plug-in called *this is your digital life* to build a targeting model.¹²² Machine learning approaches were deployed here supposedly in order to develop a pipeline through which the most easily influenced individuals could be identified–influenced to turn out to vote, to remain at home, or to change their mind on their voting preference–as well as the course of action which will make them change their view. These microtargeting models are then linked back to individuals users through identifying information such as contact details or other specific contextual information.¹²³

This debate came during two electoral results which revealed heavy political cleavages: the 2016 US election which elected President Donald Trump, and the 2016 UK referendum on exiting the European Union: both of which firm Cambridge Analytica were tied to in varying degrees. The jury remains out as to the extent to which any microtargeting practice swung any final result–and indeed, given that we cannot go back in time to run a study on this, we are likely to never know.

Several results emerged from this. Data protection regulators took regulatory action taken against Facebook and the targeting firms involved, as well as investigating these practices more widely and publishing reports on this topic.¹²⁴ In the UK, the timing of the scandal coincided with the passing of the Data Protection Act 2018 through Parliament, and created a great deal more interest in it than likely would have existed

¹²¹ See eg Carole Cadwalladr and Emma Graham-Harrison, ‘Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach’ (*The Observer*, 17th March 2018) (<https://perma.cc/HFC8-WWDT>); Carole Cadwalladr, ‘Revealed: how US billionaire helped to back Brexit’ (*The Observer*, 25th February 2017) (<https://perma.cc/PJ7A-RMN6>).

¹²² Cadwalladr, ‘Revealed: how US billionaire helped to back Brexit’ (n 121).

¹²³ See generally Giridhari Venkatadri, Elena Lucherini, Piotr Sapiezynski and Alan Mislove, ‘Investigating Sources of PII Used in Facebook’s Targeted Advertising’ [2018] Proceedings on Privacy Enhancing Technologies.

¹²⁴ See eg Information Commissioner’s Office, *Investigation into the use of data analytics in political campaigns: A report to Parliament* (, ICO 2018).

1. Hello, World!

otherwise.¹²⁵ The rapidly-passed California Consumer Privacy Act 2018,¹²⁶ which has brought some aspects of European data protection law onto US statute books for the first time, is also likely at least partially attributable to this scandal.¹²⁷

1.5. Algorithmic issues

The fraction of recent ‘war-stories’ relayed above demonstrates an array of issues concerning machine learning in society.¹²⁸

There are concerns surrounding how decisions concerning individuals reflect their distinctiveness or uniqueness as a person, or how much they are judged as part of a group. In the *COMPAS* recidivism case, there were concerns that individuals were having freedoms restored or curtailed on the basis of group membership (eg of an ethnic minority) rather than an analysis of them as an individual, on their own merits. To some, this reflects different ideas about maintaining individual human dignity in mechanised environments,¹²⁹ while for others, the focus is on how these systems affect countable outcomes that could be considered unjust. This ‘statistical discrimination’ notion¹³⁰ has been criticised on the basis that such groups seem quite fluid and arbitrary. Judging people as suitable for a position on the basis of certain characteristics (eg whether they are a smoker, whether they are physically able to safely operate certain machinery) might be acceptable, but on other characteristics, might not be. As a result, it has been forwarded that perhaps that the mechanisms for making decisions are *insufficiently precise* in their operation.¹³¹

¹²⁵ From the perspective of the author, who was involved in drafting several amendments to this Bill (see discussion in Hill, ‘Algorithms, Henry VIII powers, dodgy 1-man-firms: Reg strokes claw over Data Protection Bill’ (n 9)), the Data Protection Bill debates were in heavy danger of being swept up in debating two controversial amendments relating to the re-opening of the second part of the *Leveson* inquiry into the relationship between journalists and the police. As the Cambridge Analytica stories above (in particular Cadwalladr and Graham-Harrison (n 121)) landed in the middle of the House of Commons committee stage, the debate took a notably algorithmic turn.

¹²⁶ 2018 Cal. Legis. Serv. Ch. 55 (A.B. 375). This law is currently expected to go into effect in January 2020.

¹²⁷ Dipayan Ghosh, ‘What You Need to Know About California’s New Data Privacy Law’ (*Harvard Business Review*, 11th July 2018) (<https://perma.cc/DZ4V-AX4U>). But note that California has traditionally been a comparative leader in the US in privacy protection regardless of this act. See <https://oag.ca.gov/privacy/privacy-laws> for a list.

¹²⁸ More ‘war-stories’ can be found in AI Now Institute, *The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term* (2016) (<https://artificialintelligencenow.com/>); Campolo, Sanfilippo, Whittaker and Crawford (n 22); Michael Veale, *Data management and use: case studies of technologies and governance* (The Royal Society and the British Academy 2017); Information Commissioner’s Office, *Big data, artificial intelligence, machine learning and data protection* (ICO 2017) (<https://perma.cc/99ZT-R6TF>).

¹²⁹ Meg Leta Jones, ‘The right to a human in the loop: Political constructions of computer automation and personhood’ (2017) 47(2) *Social Studies of Science* 216 DOI: 10/f93mxz.

¹³⁰ Edmund S Phelps, ‘The Statistical Theory of Racism and Sexism’ (1972) 62(4) *The American Economic Review* 659 (<https://www.jstor.org/stable/1806107>).

¹³¹ Reuben Binns, ‘Fairness in Machine Learning: Lessons from Political Philosophy’ in *Conference on Fair-*

Yet high levels of precision also create other uneasiness around algorithmic systems. The *Cambridge Analytica* case indicated concerns around effective ‘microtargeting’ practices seemingly not correlated with any protected group, as they allowed individuals to be effectively manipulated in a highly granular way. Emerging academic concerns about the impact of optimisation systems echo this.¹³² Worries about price discrimination online share similar concerns.¹³³ Theoretically, in economics, first-degree price discrimination occurs when ‘the price exacted for [a unit of a commodity is] equal to the demand price for it’.¹³⁴ Traditionally, this has been thought of as an impossible target, as it would require a seller to know the maximum price an individual would pay for any item. Even logistically separating consumers into many different groups (as flexible and non-flexible tickets attempt to do on planes) is challenging. Separating out individuals such that they are effectively extorted for the maximum price they are willing to pay using web profiling and tracking technologies, as is increasingly commonplace,¹³⁵ seems to generate worries that instead of being insufficiently precise, systems are *overly precise* as to lose some sense of equal treatment. A comparison can be made to insurance, where the principle of equivalence, used primarily in private insurance schemes, determines an individual’s payments (to an insurance scheme) on the basis of their individual risk profile,¹³⁶ whereas the principle of solidarity, used primarily in public insurance schemes, balances the different economic situations of its members, who pay in the same proportions of their income, as they all receive the same kind of treatment.¹³⁷ This has been reflected in the ways in which many jurisdictions regulate insurance markets to avoid overly precise and granular discrimination and maintain some sense of solidarity, such as around pre-existing conditions under the US Affordable Care Act 2010, or concerning gender differences in car insurance under the *Test-Achats* CJEU case which struck down the derogation from the general rule of unisex insurance premiums and benefits in Union

ness, Accountability and Transparency (FAT 2018)* (PMLR 2018) vol 81.

¹³² Rebekah Overdorf, Bogdan Kulynych, Ero Balsa, Carmela Troncoso and Seda Gürses, ‘POTs: Protective Optimization Technologies’ [2018] arXiv preprint (<https://arxiv.org/abs/1806.02711>); Seda Gürses, Rebekah Overdorf and Ero Balsa, ‘Stirring the POTs: protective optimization technologies’ in Emre Bayamlioglu, Irina Baraliuc, Liisa Albertha Wilhelmina Janssens and Mireille Hildebrandt (eds), *BEING PROFILED: COGITAS ERGO SUM* (Amsterdam University Press 2018).

¹³³ Frederik Zuiderveen Borgesius and Joost Poort, ‘Online Price Discrimination and EU Data Privacy Law’ (2017) 40(3) *Journal of Consumer Policy* 347 DOI: 10/gdz28f.

¹³⁴ Arthur Pigou, *The Economics of Welfare* (first published 1952, Routledge 2017) DOI: 10/gfgq79 244.

¹³⁵ Zuiderveen Borgesius and Poort (n 133).

¹³⁶ ‘Principle of Equivalence’, in Wilhelm Kirch (ed), *Encyclopedia of Public Health* (Springer 2008) DOI: 10/d698f2.

¹³⁷ ‘Principle of Solidarity’, in Wilhelm Kirch (ed), *Encyclopedia of Public Health* (Springer 2008) DOI: 10/bgc69x.

1. Hello, World!

law.¹³⁸

Other issues seem to touch upon how individuals are represented in systems, even where outcomes do not flow directly from them. A duty to ensure cultural representation in official documents, public broadcasting or the like is often carried out seemingly without a need to link to any specific unequal explicit harms across groups.¹³⁹ The *Sweeney Search* case, despite not affecting Sweeney herself (although it perhaps could have done if, for example, a potential employer had searched her name), seems to indicate a problematic representation, as search results have been accused of more broadly.¹⁴⁰ Similarly, the *Word Embeddings* and *Google Gorillas* cases seem intrinsically problematic even when unlinked to a consequential decision-making system. These representational harms differ from distributive harms, yet seem to be high in public consciousness.¹⁴¹

How individuals are represented in systems or databases is also far from their control. In the *Target* war-story, an individual had data about her inferred and added to her profile without her knowledge. This reflects concerns around machine learning as transforming non-sensitive data into sensitive, and often private, insights, as did the *Cambridge Analytica* case (around political opinion or manipulability) or the *COMPAS recidivism* case (around alleged ‘danger’ or propensity to reoffend).

More broadly, there are larger concerns surrounding the circumstances that leads to any of these situations. Many of the recent computer science and statistical approaches to fairness¹⁴² attempt to rectify larger questions of inequality rooted in socioeconomics and societal power relations. Insofar as crime, gender and ethnicity are intertwined, then the *COMPAS recidivism* prediction case is always going to raise issues of where and who in the system is responsible for attempting to rectify the issue. While it seems easy to argue that a statistical model should not worsen an undesirable situation, or even retrench it, should it go further and attempt to reverse it? If so, what would the effects, such as to incentives, on other parts of the relevant system be?

Sidestepping from outcomes, there are a combined set of worries around process. The invisibility revealed in the *Cambridge Analytica* case, where individuals were concerned about targeting they could not see; the inscrutability in the *COMPAS* case, with a proprietary system that could not be inspected, or the opacity in the *Word Embeddings* case, where such biases take scientific papers to uncover, all point to difficulties for

¹³⁸ Case C-236/09 *Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des ministres* ECLI:EU:C:2011:100.

¹³⁹ Charles Taylor, *Multiculturalism: Examining the Politics of Recognition* (Princeton University Press 1994); Binns, ‘Fairness in Machine Learning: Lessons from Political Philosophy’ (n 131).

¹⁴⁰ Safiya Umoja Noble, *Algorithms of Oppression* (NYU Press 2018).

¹⁴¹ Kate Crawford, ‘The Problem with Bias’ [2017] Keynote given at NIPS 2017; Binns, ‘Fairness in Machine Learning: Lessons from Political Philosophy’ (n 131).

¹⁴² Discussed in section 1.6.2, p. 65..

individuals or organisations to grapple with potentially discriminatory or otherwise undesirable systems in and around their lives. This vision of inscrutable, out of control authority is common in earlier notions of unfathomable bureaucracy found everywhere from Kafka's *Trial* to Terry Gilliam's *Brazil*, and its appearance more broadly than visions of the public sector is indicative of the power such private processes appear to have over individual's lives in critical sectors such as information access or financial provision.

Some frameworks for an 'ethics of algorithms' do exist. I take issue with these however for a number of reasons, and consequently resist using them as an organising principle in this work. Firstly, they suffer from attempting to place all potential societal challenges with technology in a neat structure. Given the potential for algorithmic systems to be integrated in many consequential sectors and domains, this seems an effort as futile as neatly tabulating all human values. This approach has shown its thorniness most clearly in the field of development, where the arguably most influential theoretical frame in this space is the *capability approach* from development economics. This approach seeks to evaluate social arrangements based upon the extent of freedom individuals hold to promote or achieve *functionings* they value. Amartya Sen's original notion was deliberately incomplete to recognise the ways that both values change and that the framework must be operationalised differently in different contexts,¹⁴³ while later versions that sought to crystallise cross-culturally applicable 'central' human capabilities, such as those of Martha Nussbaum,¹⁴⁴ were criticised accordingly for being intellectualised rather than reflecting cross-cultural research, Western, elitist, paternalistic or inappropriate for many applications or contexts.¹⁴⁵

Turning to 'ethics of algorithms', one such framework portrays as a single category the extraordinarily broad 'unfair outcomes based on the use of the system'.¹⁴⁶ Defenders of these systems could reply by claiming that such a frame reflects only a subset of additional concerns relating to 'issues [that] apply to algorithms themselves, as op-

¹⁴³ Amartya K Sen, *Development as freedom* (Oxford University Press 1999).

¹⁴⁴ Martha C Nussbaum, *Women and human development: The capabilities approach* (Cambridge University Press 2001).

¹⁴⁵ See eg Nivedita Menon, 'Universalism without Foundations?' (2002) 31(1) *Economy and Society* 152 DOI: 10/fsc89q; Susan Moller Okin, 'Poverty, Well-Being, and Gender: What Counts, Who's Heard?' (2005) 31(3) *Philosophy & Public Affairs* 280 DOI: 10/cr3hq; Sabina Alkire, *Valuing freedoms: Sen's capability approach and poverty reduction* (Oxford University Press 2005).

¹⁴⁶ Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter and Luciano Floridi, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3(2) *Big Data & Society* 2053951716679679 DOI: 10/gcdx92. These sit among the inconclusive and uncertain nature of the evidence being used; the inscrutable nature of the connection between data and the conclusion drawn; misguided evidence calling the neutrality of the process into question; unfair outcomes based on the use of the system; transformative effects of systems on social and political organisation; and the difficulty that these system cause for the traceability of ethical responsibility.

1. Hello, World!

posed to technologies built upon algorithms'.¹⁴⁷ Yet this too assumes such a division is either possible or useful to make for anything other than intellectual purposes. The services uniquely enabled by algorithmic systems (eg real time content moderation at scale) cannot be meaningfully separated from the algorithmic systems themselves whilst still discussing the issues of societal concern. Research designed to inform policy risks, if it is not careful, promotion the confusion of such a conceptual framework for a complete guide to the impacts society expects regulation of.

Consequently, I believe an parsimonious and deployable 'ethics of algorithms' must necessarily be so narrowly coupled to the software itself to either be primarily of only academic interest by missing out genuine social and policy challenges linked to algorithmic systems more broadly, or so expansive as to blur with broader ethical frameworks and be unconnected to algorithmic systems in many ways beyond branding. This does not mean that individual issues cannot be defined and approached by ethicists in context, but simply an argument against the utility of universal organising principles in this expansive domain. Indeed, I personally feel that the main beneficiaries of the impossibility of universality are those organisations and institutions who wish to keep certain aspects of algorithmic systems that might threaten their interests *out* of debate, rather than those individuals seeking to include and structure issues in a relatively comprehensive way.¹⁴⁸ Consequently, it an approach I wish to avoid in this research. Defining problems and envisioning a world without them (or with fewer of them) is messy, and this messiness is not to be shied away from.¹⁴⁹

1.6. Computing to the Rescue?

Technical fields doing research with downstream social impacts are often accused of having little awareness or consideration of their work's potential consequences. It could be said that this is not a bad state of affairs: that the creation of knowledge by scientists is a good which trumps other forms of responsibility,¹⁵⁰ that technological development can be so unpredictable as to render prudent calculation of adverse social impacts as futile,¹⁵¹ or that innovation systems involve so many interacting actors

¹⁴⁷ Mittelstadt, Allo, Taddeo, Wachter and Floridi (n 146) 2.

¹⁴⁸ See eg Ben Wagner, 'Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping?' in Emre Bayamloğlu, Irina Baraliuc, Liisa Albertha Wilhelmina Janssens and Mireille Hildebrandt (eds), *BEING PROFILED: COGITAS ERGO SUM* (Amsterdam University Press 2018).

¹⁴⁹ See generally Robert Hoppe, *The Governance of Problems: Puzzling, Powering and Participation* (Policy Press 2010).

¹⁵⁰ Heather E Douglas, 'The Moral Responsibilities of Scientists (Tensions between Autonomy and Responsibility)' (2003) 40(1) *American Philosophical Quarterly* 59.

¹⁵¹ See generally the debates around the problem of 'moral luck' and subsequent responsibility. BAO Williams and T Nagel, 'Moral Luck' (1976) 50 *Proceedings of the Aristotelian Society*, Supplementary

that any individualised notion of responsibility significantly overemphasises the role of any single actor.¹⁵² Initiatives pushing back against these arguments have primarily done so by sidestepping individual responsibility or capacity in favour of mechanisms for establishing more pluralist, forward looking forms of responsibility, inclusion, steer and oversight.¹⁵³

In machine learning, the approach to science governance has looked a little different from this vision. Instead of directing the development of technologies with societal steer, the response that computer scientists have engaged in (as opposed to critical scholars from other disciplines) has been to build *further* technologies designed to ameliorate shortcomings of the initial tools. Such shortcomings have been framed to primarily concern issues of fairness and non discrimination; issues of transparency and scrutability, and issues of accountability. This can be understood in the context of similar technological developments. These often come with grandiose promises of an imagined future, and the imaginaries shared by a research community are often responded to in further work. Applied machine learning researchers intend to optimise tasks those developing it or sponsoring its development consider important. The idea that optimisation of societal functions *is* a good thing is rarely questioned.¹⁵⁴ Instead, attention is turned to whatever might threaten machine learning from succeeding in this task, according to some notion of success. Similar technologies with such *intent* have also been the subject to technical fixes.¹⁵⁵ Gene drives for instance, which might propagate genetically modified traits across entire populations, usually intending to eg reduce malaria prevalence among mosquitos, have been the subject of proposed new augmentations that might manage their risk or mitigate dangerously transformative effects in the wild (by, for example, ensuring that their transformative potential deteriorates over generations).¹⁵⁶ In climate engineering (also known as geoengineering), which concerns a panoply of hypothetical technologies intended forestall the impacts of global warming, researchers consider the threats to its success more than they question its aims in general.¹⁵⁷

Volumes 115 DOI: 10/gfphtz.

¹⁵² René von Schomberg, *From the ethics of technology towards an ethics of knowledge policy & knowledge assessment* (European Commission 2007) (<https://perma.cc/Q7KQ-CQLE>).

¹⁵³ See generally Jack Stilgoe, Richard Owen and Phil Macnaghten, 'Developing a framework for responsible innovation' (2013) 42(9) *Research Policy* 1568 DOI: 10/f5gv8h.

¹⁵⁴ For an account that does criticise optimisation, see Overdorf, Kulynych, Balsa, Troncoso and Gürses (n 132); Gürses, Overdorf and Balsa (n 132).

¹⁵⁵ cf Jack Stilgoe, *Experiment Earth: Responsible Innovation in Geoengineering* (Earthscan 2015) in relation to climate engineering and intent.

¹⁵⁶ See Kenneth A Oye and others, 'Regulating Gene Drives' (2014) 345(6197) *Science* 626 DOI: 10/gfphx3; Kevin M Esvelt, Andrea L Smidler, Flaminia Catteruccia and George M Church, 'Concerning RNA-Guided Gene Drives for the Alteration of Wild Populations' (2014) 3 *eLife* DOI: 10/gfphx6

¹⁵⁷ See generally Stilgoe (n 155).

1. Hello, World!

What are the fixes designed to ensure that the predestined aim of optimisation by machine learning is not threatened or derailed? I will now outline three groups of fixes that have fixated the machine learning community alongside their common monikers. These are techniques designed to increase transparency (Explaining), to increase fairness (Debiasing), and to increase accountability (Accounting).

1.6.1. Explaining

Explanations in computer systems are far from a new topic of study or practice. As this section will discuss, and contrary to some recent work, not only have they been a topic of research for some decades, but many important distinctions have already emerged and been elaborated upon.

Calls for systems to explain themselves have a recent resurgence in the algorithmic war stories described above.¹⁵⁸ Stories such as *COMPAS*¹⁵⁹ provoke visions of out-of-control automated justice with lacking justification or recourse. While these systems might not always themselves be so complex as to resist all understanding on careful inspection, the lack of understanding can jointly stem from the protection software systems are afforded by the intellectual property regime, particularly the law of trade secrecy.¹⁶⁰

In relation to these recent concerns, the computer science community has responded with the field of ‘explainable AI’, occasionally called XAI. The top machine learning conferences seen as necessary for academic promotion such as Neural Information Processing Systems (NeurIPS) or the International Conference on Machine Learning (ICML) have recently initiated explicit paper streams or categories for accountable and/or explainable AI, and the credibility and academic kudos for research in this field has seen a sharp increase accordingly. This has been further spurred by a range of explicit grant calls in the area from organisations such as the Defense Advanced Research Projects Agency (DARPA) or the Engineering and Physical Sciences Research Council (EPSRC).¹⁶¹

As a general principle, it has however *long* been held that systems should be able account for their own operation not as an end in and of itself, but in ways that help their users understand how their tasks are being accomplished.¹⁶² Explanations seem par-

¹⁵⁸ Section 1.4, p. 39.

¹⁵⁹ Section 1.4.1, p. 40.

¹⁶⁰ See generally Frank Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information* (Harvard University Press 2015).

¹⁶¹ See eg the Trust, Identity, Privacy and Security funding stream for EPSRC, and the XAI programme for DARPA.

¹⁶² Paul Dourish, ‘Accounting for system behaviour: Representation, reflection and resourceful action’ [1997] *Computers and Design in Context* 145; Ben Shneiderman and Pattie Maes, ‘Direct manipulation

ticularly useful in building confidence and acceptance¹⁶³—for example, insofar as they help a user discover when the heuristics underpinning a system fail in practice, or a narrowly conceived system is being ‘pushed beyond the boundaries of its expertise’.¹⁶⁴ Some work even argues humans are inherently unlikely to trust machine decisions even when they are statistically sound,¹⁶⁵ and that this is problematic where expert systems can provide useful assistance in complex tasks that may outperform humans in certain respects, or help overcome specific cognitive biases.¹⁶⁶ When systems are safety-critical, understanding them in order to make informed decisions about their deployment and management is subsequently of great concern.¹⁶⁷

In the heady days of expert systems, it was believed that specialist knowledge such as that of a clinician could be represented in software, often through a body of manually-coded if-then rules obtained through knowledge elicitation techniques.¹⁶⁸ Explanations mattered here, too, and early knowledge representation and reasoning systems typically aimed to strive for trust by producing explanations alongside their outputs. One of the most famous early expert systems, MYCIN,¹⁶⁹ designed to provide advice regarding selection of appropriate antimicrobial therapy for hospital patients with bacterial infections, sought to explain itself so by answering questions about ‘why’ it was aiming at a current goal, and ‘how’ it was achieving it. These explanations, as well as those of other contemporary systems, were based on only the information used in the delivery of the expert system itself, such as explanations derived from the rules in rule-based architectures.¹⁷⁰ Often called *explanation facilities*, these were further augmented by explicitly modeling the knowledge engineering and

vs. interface agents’ (1997) 4(6) *Interactions* 42.

¹⁶³ LRichard Ye and Paul E Johnson, ‘The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice’ (1995) 19(2) *MIS Quarterly* 157 DOI: 10/brgtbj.

¹⁶⁴ William R Swartout, ‘XPLAIN: A system for creating and explaining expert consulting programs’ (1983) 21(3) *Artificial Intelligence* 285 DOI: 10/bxgmhx, 286.

¹⁶⁵ For an example of a study with such a logic, see Berkeley J Dietvorst, Joseph P Simmons and Cade Massey, ‘Algorithm aversion: People erroneously avoid algorithms after seeing them err’ (2015) 144(1) *Journal of Experimental Psychology: General* 114 DOI: 10/f6xqfw.

¹⁶⁶ See eg Amos Tversky and Daniel Kahneman, ‘Judgment under Uncertainty: Heuristics and Biases’ (1974) 185(4157) *Science* 1124.

¹⁶⁷ Włodzisław Duch, ‘Coloring black boxes: visualization of neural network decisions’ (2003) 3 *International Joint Conference on Neural Networks* 1735.

¹⁶⁸ Such as Nancy J Cooke, ‘Varieties of knowledge elicitation techniques’ (1994) 41(6) *International Journal of Human-Computer Studies* 801 DOI: 10.1006/ijhc.1994.1083; Robert R Hoffman, Beth Crandall and Nigel Shadbolt, ‘Use of the critical decision method to elicit expert knowledge: A case study in the methodology of cognitive task analysis’ (1998) 40(2) *Human Factors* 254 DOI: 10.1518/001872098779480442.

¹⁶⁹ See generally Edward Shortliffe, *Computer-Based Medical Consultations: MYCIN* (Elsevier 1976).

¹⁷⁰ Randy L Teach and Edward H Shortliffe, ‘An analysis of physician attitudes regarding computer-based clinical consultation systems’ (1981) 14(6) *Computers and Biomedical Research* 542; William J Clancey, ‘The epistemology of a rule-based expert system—a framework for explanation’ (1983) 20(3) *Artificial Intelligence* 215.

1. Hello, World!

design processes and revealing information about them to decision-makers,¹⁷¹ as well as providing extra information beyond the ‘trace’ of the software execution that would contextualise the decision or advice.¹⁷² Inspired by different theoretical models of argumentation, various explanation types have been explored and tested, with the style of explanation provided affecting the extent to which decision-makers understand the decisions of expert systems.¹⁷³

As expert systems began to integrate more statistical components as opposed to bodies of if-then rules, their explanation facilities had to leave primarily rule-based paradigm and venture into murkier, more explicitly probabilistic waters.¹⁷⁴ This accompanied the move from expert systems attempting to mimic the reasoning of human experts, using structured objects and rule systems to build parallels to experts’ heuristics, to systems attempting to model whole systems in ways that might not match an expert’s worldview, using approaches such as Bayesian networks.¹⁷⁵ While the latter approach often led to ‘more accurate and robust expert systems’, it also made explanation facilities ‘even more necessary, because normative reasoning methods are more foreign to human beings than heuristic methods.’¹⁷⁶

With a few exceptions,¹⁷⁷ explanation facilities in the expert systems literature have historically primarily been linked to those *administering* decisions, rather than those on the receiving end of them. This placed the focus squarely on how one could account for the users’ knowledge about the domain, and design explanation facilities best suited to enabling some task.¹⁷⁸ There are several possible reasons for this.

Firstly, there was a constant difficulty in moving expert systems from research to

¹⁷¹ Robert Neches, William R Swartout and Johanna D Moore, ‘Enhanced maintenance and explanation of expert systems through explicit models of their development’ [1985] (11) IEEE Transactions on Software Engineering 1337.

¹⁷² Michael R Wick and William B Thompson, ‘Reconstructive expert system explanation’ (1992) 54(1-2) Artificial Intelligence 33 DOI: 10/d529gf.

¹⁷³ Johanna D Moore and William R Swartout, ‘A reactive approach to explanation: Taking the user’s feedback into account’ in Cecile Paris, William R Swartout and William C Mann (eds), *Natural Language Generation in Artificial Intelligence and Computational Linguistics* (Springer 1991); Brian Y Lim, Anind K Dey and Daniel Avrahami, ‘Why and why not explanations improve the intelligibility of context-aware intelligent systems’ in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI’09)* (ACM 2009) DOI: 10/cq2626.

¹⁷⁴ See eg Suran Goonatilake and Sukhdev Khebbal (eds), *Intelligent Hybrid Systems* (John Wiley & Sons, Inc 1994).

¹⁷⁵ Carmen Lacave and Francisco J Díez, ‘A review of explanation methods for Bayesian networks’ (2002) 17(2) The Knowledge Engineering Review 107.

¹⁷⁶ *ibid* 39.

¹⁷⁷ While broadly not utilising statistical explanation systems, there are examples of interdisciplinary projects in the expert systems domain aiming explanations and those subject to decisions, rather than administering them. See eg Diana E Forsythe, ‘Using ethnography in the design of an explanation system’ (1995) 8(4) Expert Systems with Applications 403 DOI: 10/c924fs; Diana E Forsythe, ‘New bottles, old wine: Hidden cultural assumptions in a computerized explanation system for migraine sufferers’ (1996) 10(4) Medical Anthropology Quarterly 551 DOI: 10/b4hp3p.

¹⁷⁸ Lacave and Díez (n 175) 39.

deployment, adoption and wider dissemination.¹⁷⁹ Even one of the most famous systems, MYCIN, did not see use in practice. In these cases, the acceptability of decisions by decision-support users was seen as more important than the acceptability of those decisions to decision subjects, as without practitioners appreciating, exploring the utility of, and deploying these expert systems, the last question seemed a distant challenge.

Secondly, expert systems were designed either to turn lay users into specialists or to heighten the specialism (or decrease cognitive load) on more general experts or professionals. They were not designed to create a seamless or fully automated process (although this may have been the long-term vision of some of their key architects). As a result, lay operators could serve as an intermediary to parse an explanation as *they* understood it into a form that the decision subject could understand. Furthermore, where such a lay user was expected to augment the decision with their own knowledge of the context in which it was being made in, any explanation aimed at a decision-subject would need to incorporate both the system output *and* its subsequent augmentation, meaning that explanation facilities not aimed at the decision-maker might often produce inappropriate or confusing results, placing further interpretive load on decision-subjects (and coming at further detriment to the ‘human factor’ which was likely desirable for other reasons, such as trust or comfort).

This does not mean no parallels to decision-subject facing work being carried out today existed then. Perhaps the closest work of this flavour can be found in the study of knowledge/rule extraction. This field aimed to use the inferential power of a statistical system, such as a neural network, to get at some underlying, human-understandable insight about a domain.¹⁸⁰ Often, this took the form of attempting to extract and refine interpretable ‘rules’ from a system, which could be then be parsed to better understand phenomena of interest. These rules would then be used in practice, rather than the predictive model that had been first generated. Indeed, some studies of rule extraction have cited their motivation as explanation requirements in law,¹⁸¹ such as the US Fair Credit Reporting Act (FCRA)¹⁸² and the Equal Credit Opportun-

¹⁷⁹ Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey and James F Antaki, ‘Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help’ in *Proceedings of the 2016 SIGCHI Conference on Human Factors in Computing Systems (CHI’16)* (ACM 2016) DOI: 10/gddkjt.

¹⁸⁰ Robert Andrews, Joachim Diederich and Alan B Tickle, ‘Survey and critique of techniques for extracting rules from trained artificial neural networks’ (1995) 8(6) *Knowledge-Based Systems* 373; Alan B Tickle, Robert Andrews, Mostefa Golea and Joachim Diederich, ‘The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks’ (1998) 9(6) *IEEE Transactions on Neural Networks* 1057 DOI: 10/btn5vv.

¹⁸¹ David Martens, Bart Baesens, Tony Van Gestel and Jan Vanthienen, ‘Comprehensible credit scoring models using rule extraction from support vector machines’ (2007) 183(3) *European Journal of Operational Research* 1466.

¹⁸² 15 U.S.C. §1681, *et seq.*

1. Hello, World!

ity Act (ECOA)¹⁸³, both of which require ‘adverse action notices’ to contain a statement of reasons for any denial of credit based on the scoring and aggregation systems used.¹⁸⁴ The research motivation here was to create ‘comprehensible classification techniques’, comparable in their generalisable performance to ‘black-boxed’ approaches such as support vector machines or neural networks, and thus legally permitted in these contexts with few knock-on effects on accuracy or efficacy.¹⁸⁵

Another domain where explanations of automated systems aimed at users can be found is within HCI. In context-aware computing, a range of explanations have been tested with users in areas such as wearables for health detection and prediction or ubiquitous computing devices. Such explanations often focussed on ‘why’ and ‘why not’ systems behaved or predicted user activity in certain ways, and what changes users could have made in input data to make the output otherwise.¹⁸⁶ Research in this domain focussed on styles of explanation rather than methods for extracting such explanation from statistical systems, which were often not involved in the applications in question. In the more statistically oriented fields of recommender systems and collaborative filtering, an array of explanation types have also been deployed, primarily seeking to improve user experience.¹⁸⁷

These strands of research have become more relevant of late. While the need for explanations for users of decision-support has not gone away, explanations and machine learning systems in popular consciousness are now much more intimately linked to the ideal of effective control of algorithmic decision-making in relation to those affected by decisions, rather than those making them. Tal Zarsky argues that individuals adversely affected by a predictive process have the right to ‘understand why’, and frames this in familiar terms of autonomy and respect as a human being.¹⁸⁸ Mireille Hildebrandt has long called for transparency-enhancing tools to control the impacts of profiling technologies.¹⁸⁹ Metaphors such as the ‘black box’ have been further popularised by accessible works on algorithmic systems by authors such as Frank Pasquale and Cathy O’Neill.¹⁹⁰ These have been reinforced by a staggering array of

¹⁸³ *ibid* §1691, *et seq.*

¹⁸⁴ Andrew Selbst and Solon Barocas, ‘The Intuitive Appeal of Explainable Machines’ (2018) 87 *Fordham L. Rev.* 1085 DOI: 10/gdz285.

¹⁸⁵ Martens, Baesens, Van Gestel and Vanthienen (n 181).

¹⁸⁶ Lim, Dey and Avrahami (n 173); Brian Y Lim and Anind K Dey, ‘Assessing Demand for Intelligibility in Context-Aware Applications’ in *Proceedings of the 11th International Conference on Ubiquitous Computing (UbiComp ’09, ACM 2009)* DOI: 10/dtkpgv.

¹⁸⁷ See generally Nava Tintarev and Judith Masthoff, ‘Designing and evaluating explanations for recommender systems’ in *Recommender Systems Handbook* (Springer 2011).

¹⁸⁸ Tal Zarsky, ‘Transparency in data mining: From theory to practice’ in Bart Custers, Toon Calders, Bart Schermer and Tal Zarsky (eds), *Discrimination and privacy in the information society* (Springer 2013) 317.

¹⁸⁹ Mireille Hildebrandt, ‘Profiling and the rule of law’ (2008) 1(1) *Identity in the Information Society* 55 DOI: 10.1007/s12394-008-0003-1.

¹⁹⁰ Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information* (n 160); Cathy

policy reports (with a considerable degree of redundancy) that have emphasised explainability to decision-subjects as one of the core challenges concerning algorithmic systems.¹⁹¹

This change in winds has occurred to the point that much of the literature referred to above—both decision-subject and decision-maker facing—seems forgotten in many contemporary discussions.¹⁹²

The clearest strand that remains builds on the ‘comprehensible classification techniques’ described above.¹⁹³ A narrative has crystallised in many fields, driven primarily by the hype around machine learning outlined in the introduction, that there is utility to be had from more complex models in many sensitive contexts. Deep learning models in particular, where feature constructs become difficult for humans to parse due to being optimised for computational effectiveness rather than semantic parsability, suffer from this.¹⁹⁴ Yet this narrative deserves further scrutiny.

A field of research has pushed back on ‘explaining’ complex machine learning systems to argue instead that it is possible to make simpler but comparatively performant systems by working harder on the ways that these systems are being optimised. Especially in domains such as recidivism prediction and healthcare, interpretable formulae or scorecards appear to rival methods such as random forests, support vector machines or neural networks.¹⁹⁵ These approaches struggle in high dimensional domains, which are usually characterised by a great number of input variables lacking intrinsic human meaning (such as pixels, or sensor readings), yet their proponents

O’Neil, *Weapons of Math Destruction* (Penguin 2016).

¹⁹¹ See eg Science and Technology Committee (Commons), *Algorithms in Decision-Making* (HC 2018, 351) (<https://perma.cc/PH52-NUWC>) 27; Select Committee on Artificial Intelligence (Lords), *AI in the UK: ready, willing and able?* (HL 2018, Paper) 36; Wetenschappelijke Raad voor het Regeringsbeleid, *Big Data in een vrije en veilige samenleving (WRR-Rapport 95)* (WRR 2016) (<http://www.wrr.nl/publicaties/publicatie/article/big-data-in-een-vrije-en-veilige-samenleving/>); National Science and Technology Council, *Preparing for the future of artificial intelligence* (US Government 2016) (<https://perma.cc/DDR3-2QBH>) 30; *Artificial Intelligence—a European perspective* (n 4); Committee of experts on internet intermediaries (MSI-NET) (n 3) 36; The Royal Society (n 18); Royal United Services Institute for Defence and Security Studies (RUSI) (n 7); Committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT) (n 3); Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (n 3); Information Commissioner’s Office, *Big data, artificial intelligence, machine learning and data protection* (n 128), among many others.

¹⁹² For a recent survey considering this literature and trying to reintroduce it into the contemporary discussion, see Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi and Fosca Giannotti, ‘A Survey Of Methods For Explaining Black Box Models’ [2018] arXiv preprint (<https://arxiv.org/abs/1802.01933v>).

¹⁹³ Martens, Baesens, Van Gestel and Vanthienen (n 181).

¹⁹⁴ Jenna Burrell, ‘How the machine ‘thinks’: Understanding opacity in machine learning algorithms’ (2016) 3(1) *Big Data & Society* DOI: 10.1177/2053951715622512.

¹⁹⁵ Jiaming Zeng, Berk Ustun and Cynthia Rudin, ‘Interpretable classification models for recidivism prediction’ (2017) 180(3) *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 689 DOI: 10.1111/rssa.12227; Berk Ustun and Cynthia Rudin, ‘Supersparse Linear Integer Models for Optimized Medical Scoring Systems’ (2016) 102(3) *Machine Learning* 349 DOI: 10/f8crhw.

1. Hello, World!

emphasise that there *are* many cases, particularly those where the stakes are high, which are highly amenable to this type of interpretable modelling.¹⁹⁶ It is important to note that many models that have been subject to the popular war stories described might themselves *not* be so complex as to resist these comprehensible classification approaches. The company responsible for creating the COMPAS tool has stated, for example, that only 6 input factors go into its model.¹⁹⁷

Broadly, these attempts to make more interpretable systems *ab initio* are aimed at the model as a whole. Interpretable systems are supposed to be interpretable throughout, and can be grasped by an individual seeking to understand their functioning more generally.¹⁹⁸ Explanations of this type can be thought of as falling within a broader category of *model-centric* explanations.¹⁹⁹ In some cases, these might fall solely on characteristics of the model software itself. Linear regression for example, being linear, exhibits the same coefficients throughout the model, which is entirely defined by a relatively simple equation that can be significantly interrogated by eyeballing it. Consequently, these coefficients are widely used in domains such as quantitative political science to attempt to explain phenomena, rather than just predict them.²⁰⁰ Some authors have called this ‘system functionality’.²⁰¹ Yet this is a heavily limiting notion of what it might be to explain at the level of an entire model.²⁰²

Model-centric explanations might also offer process-related information about how a model was trained and evaluated. On which data-sources was it built? Through which testing methods was it examined, and what are the relevant performance metrics of summary statistics produced? Was it subject to simulations or in-the-wild tests with users? Were issues such as those already discussed in this thesis, such as fairness, examined for, and if so, how? Is it regularly retrained or overseen? What did those affected think of models, through processes such as user testing or through consultation? Interpretable models themselves can be seen as falling within a subset of efforts to understand global properties and characteristics of deployed systems.

In contrast, a range of techniques attempt to make explanations that are a function

¹⁹⁶ Cynthia Rudin, ‘Please Stop Explaining Black Box Models for High Stakes Decisions’ in *NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning* (2018) (<https://arxiv.org/abs/1811.10154>).

¹⁹⁷ Northpointe, Incd/b/a equivant, ‘Official Response to Science Advances’ (*Equivant (press release)*, 17th January 2018) (<http://perma.cc/YB6F-PZW9>).

¹⁹⁸ It is worth noting however that interpretable systems often do not consider the act of interpretation itself as important, and rarely ask the question of what the model is being interpreted for.

¹⁹⁹ See Edwards and Veale, ‘Slave to the Algorithm?’ (n 79) 54.

²⁰⁰ See eg Gary King, Robert O Keohane and Sidney Verba, *Designing social inquiry: Scientific inference in qualitative research* (Princeton University Press 1994).

²⁰¹ 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 DOI: 10/gfc7bb.

²⁰² See further Edwards and Veale, ‘Slave to the Algorithm?’ (n 79); Selbst and Barocas (n 184).

of an individual decision-subject or decision made about her. Such explanations are *subject-centric* insofar as they are local to a particular individual or input vector. These can be provided both after a decision is made, or in advance of it by querying an explanation facility with a hypothetical.²⁰³ A variety of flavours of explanation of this type can be derived. These include

Sensitivity-based subject-centric explanations. These ask ‘what changes in input data would have made a decision turn out otherwise?’ In a classifier, they are considering what it would take to ‘flip’ a classification to another, while in a regressor, they would consider what would provoke a change large enough to trigger a threshold. Some approaches take a soft boundary and attempt to summarise the relative importance or influence of different changes,²⁰⁴ but such a reduction is challenged by the difficulty of dealing with the importance of selected interaction effects between variables. Others try to produce one or more alternative positions with different classifications to display to end users,²⁰⁵ but are plagued by selecting cases that are possible, relevant or feasible for users from the huge number possible in the high dimensional spaces many classifiers work in, such as those that do not require users to change aspects about them which they cannot (such as their age).²⁰⁶ All sensitivity-based explanations also suffer when the variables that enter a model are not humanly interpretable. It is impossible to talk about what a decision-subject should or could change if the input data is abstract, such as the gyrometer readings from a mobile phone over several weeks, or traces of geospatial data, or telemetry data from app usage. In certain rare cases, successful ways to render high dimensional abstract data into a human-understandable form do exist. Pixels form images, which our brains can reason about. Yet this remains the clear exception, rather than the rule.

Saliency-based subject-centric explanations. These ask which features are *important* for a prediction.²⁰⁷ To do so, they usually perturb data inputs repeatedly. Primarily

²⁰³ cf Wachter, Mittelstadt and Floridi (n 201), who conflate model querying with final decision-making. The same model queried twice with the same input data will always produce the same output (assuming the random seed is also preserved, as is standard practice in reproducible research, if one is used to eg add noise).

²⁰⁴ See eg Anupam Datta, Shayak Sen and Yair Zick, ‘Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems’ in *2016 IEEE Symposium on Security and Privacy (SP)* (2016).

²⁰⁵ Sandra Wachter, Brent Mittelstadt and Chris Russell, ‘Counterfactual explanations without opening the black box: Automated decisions and the GDPR’ [2018] Harv. J.L. & Tech.

²⁰⁶ Berk Ustun, Alexander Spangher and Yang Liu, ‘Actionable Recourse in Linear Classification’ in *Proceedings of the ACM Conference on Fairness, Accountability and Transparency (ACM FAT* 2019)* (ACM 2019) (<http://arxiv.org/abs/1809.06514>).

²⁰⁷ Examples include Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek and Klaus-Robert Müller, ‘Explaining nonlinear classification decisions with deep Taylor decomposition’

1. Hello, World!

these are useful in their inverse: which parts of a data vector are *not* being considered by a model. For example, the saliency-based explanation system LIME²⁰⁸ was applied to a predictor which appeared to successfully separate images of huskies from images of wolves, despite them looking relatively similar. The explanation facility highlighted that the salient area was the background of the image while the animals themselves were generally omitted—the system appeared to instead be detecting the snow in the husky photos, rather than ‘looking’ at the animals. The human leap at logic at the end was key to the real explanation. Yet where parts of (eg an image) are considered salient, saliency-based explanations say little-to-nothing about how these parts of the data are being used: just that they are. Such systems often fail to hypothesise about what might be otherwise in the areas they are not considering due to the limits of their perturbation approaches. An image of a woman standing in a relatively plain field might be classified simply as a woman with little extra information. But if the perturbation method had added a combine harvester into the field, she might have been classified as a farmer: so how far can we say that the contents of the field are truly not salient?

Case-based subject-centric explanations. These usually focus on making comparisons between a decision-subject and historic data or course of action which might provide insight as to why a model made the choices it did. Machine learning systems are in essence *similarity engines*, with outputs based on similar cases in the training set which the system has determined are likely to generalise well. The *nearest neighbour* algorithm is the clearest example of this: assuming that the nearest (according to some distance metric) point (or average of k nearest points) will predict the point queried with. These types of explanations have been popular in expert systems²⁰⁹ but less emphasised in machine learning systems, which generally do not retain original datasets alongside the model, and to do so would potentially create additional privacy concerns not currently present.²¹⁰

Demographic-based subject-centric explanations. These ask ‘what are the characteristics of individuals who received similar treatment to the individual whose data is

(2017) 65 Pattern Recognition 211 DOI: 10/f9vv35; Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin, “Why should I trust you?": Explaining the predictions of any classifier' in (2016) DOI: 10/gfgrbd. Note that Ribeiro et al also contains a notion of directionality in comparison to a particular classification, and so could be seen as a hybrid salience–sensitivity explanation.

²⁰⁸ Ribeiro, Singh and Guestrin, “Why should I trust you?": Explaining the predictions of any classifier' (n 207).

²⁰⁹ Dónal Doyle, Alexey Tsymbal and Pádraig Cunningham, *A Review of Explanation and Explanation in Case-Based Reasoning* (, Department of Computer Science, Trinity College, Dublin 2003).

²¹⁰ See further section 3.2.

being queried?’²¹¹ While they act in reference to an individual decision-subject, they utilise data about how others are treated in the systems. For example, if a user did not qualify for a cheap tier of car insurance, they could be told the percentage of users who did; the percentage of users who had been in one accident which was not their fault; the percentage of users who regularly travelled at night, or so on.²¹²

Performance-based subject-centric explanations. It is also possible to imagine explanations aimed at individuals based on the certainty of particular queries. How confident is the system of the outcome? Are individuals making these type of queries disproportionately likely to be classified erroneously more or less the average? These ‘explanations’ focus primarily on confidence rather than determining why a result was what it was. Communicating uncertainty is a considerable field which has been applied to user interfaces at an individual level before,²¹³ although this is not usually framed as an explanation, I argue it can construe (part) of one.

The distinction between model-centric and subject-centric explanations proposed here²¹⁴ is similar but importantly not identical to the distinction between *global* and *local* explanations that has often been drawn in the explanation literature.²¹⁵ Subject-centric explanations differ from (but incorporate) local explanations. Local explanations attempt to simplify the extent of what needs to be explained by only presenting analysis based on a decision-space proximate to the decision-subject’s input vector. In doing so, they attempt to reduce the difficulty of explaining non-linearities or non-monotonicity that might permeate the entire model and lead it to resist faithful, meaningful or parseable summarisation. They assume that while it might not be possible to make a human-understandable explanation while maintaining *global fidelity*, it might be possible to do when only trying to preserve *local fidelity*.²¹⁶ Yet subject-centric

²¹¹ See eg Liliana Ardissono, Anna Goy, Giovanna Petrone, Marino Segnan and Pietro Torasso, ‘Intrigue: Personalized recommendation of tourist attractions for desktop and hand held devices’ (2003) 17(8-9) Applied Artificial Intelligence 687.

²¹² Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao and Nigel Shadbolt, ‘It’s Reducing a Human Being to a Percentage’; Perceptions of Justice in Algorithmic Decisions’ in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI’18)* (ACM 2018) DOI: 10/cvcp.

²¹³ See eg Miriam Greis, Jessica Hullman, Michael Correll, Matthew Kay and Orit Shaer, ‘Designing for Uncertainty in HCI: When Does Uncertainty Help?’ in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA ’17, ACM 2017) DOI: 10/gfn45s; Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman and Matthew Kay, ‘Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making’ in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI ’18, ACM 2018) DOI: 10/gfrfc; Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale and Matthew Kay, ‘In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation’ [2018] IEEE Transactions on Visualization and Computer Graphics DOI: 10/gfn45q.

²¹⁴ First proposed by the author in Edwards and Veale, ‘Slave to the Algorithm?’ (n 79).

²¹⁵ See eg distinctions made by Finale Doshi-Velez and others, ‘Accountability of AI under the law: The role of explanation’ [2017] arXiv preprint (<https://arxiv.org/abs/1711.01134>); Wachter, Mittelstadt and Floridi (n 201).

²¹⁶ Ribeiro, Singh and Guestrin, ‘Why should I trust you?’: Explaining the predictions of any classifier’

1. Hello, World!

explanations are simply those that relate to an individual decision, which might be across the entire model. For example, demographic explanations are not restrained to a limited distance around a decision-subject in the model, but consider the model behaviour globally, but in *relation* to a particular data point. To this extent, they highlight the finding that individuals appreciate explanations that relate to some part of their identity, rather than those that might, for example, prioritise faithfulness or fidelity to the model.²¹⁷

Explanation facilities aimed at saying something about the decision process of the model (eg sensitivity or saliency explanations as opposed to performance or demographic explanations) require some link to the original model. The nature of this link can vary. While original expert system explanations attempted to just make the same reasoning the software used explicit to the user, it was noted that the task of explanation often had different aims than the task of prediction, and as a consequence, different models were likely required.²¹⁸ Two main approaches to generating these models were proposed.²¹⁹ *Decompositional* approaches sought to use structural information from the model's innards in order to inform a model that was suitable for building explanations. *Pedagogical* approaches instead sought to solely use the model as an oracle taking input vectors and producing output vectors, and thorough examination of this 'black box', build an explanation system or extract rules. These have also been described by more recent authors as 'model-agnostic' methods.²²⁰ This distinction has been occasionally lost and confused, but remains important when considering who can make explanation systems, and what level of access they need.²²¹

How do users feel about these varied types of subject-centric explanations? I considered this question with colleagues in a paper produced during the course of this thesis, examining how different types of explanations from the categories above, presented to users (in labs and online) during scenarios inspired by war stories such as those already outlined, such as being refused insurance, or a loan, or a promotion, affected their notion of procedural justice during the process in question.²²² Firstly, it

(n 207).

²¹⁷ Motahhare Eslami, Sneha R Krishna Kumaran, Christian Sandvig and Karrie Karahalios, 'Communicating Algorithmic Process in Online Behavioral Advertising' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM 2018) DOI: 10/cxrf.

²¹⁸ Wick and Thompson (n 172).

²¹⁹ Tickle, Andrews, Golea and Diederich (n 180).

²²⁰ Ribeiro, Singh and Guestrin, "Why should I trust you?": Explaining the predictions of any classifier' (n 207); Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin, 'Model-Agnostic Interpretability of Machine Learning' in *Presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY (2016) (<https://arxiv.org/abs/1606.05386>).

²²¹ See eg Wachter, Mittelstadt and Russell (n 205), who claim their method does not '[open] the black box', but which would require internal model gradients in order to calculate, thus resembling a *decompositional* approach.

²²² Binns, Van Kleek, Veale, Lyngs, Zhao and Shadbolt (n 212).

appeared that the type of explanation provided was not a large factor in affecting notions of justice compared to provision of no explanation. Users presented with different types of explanations in the same session showed some evidence of distinct preferences between them, disliking case-based explanations and favouring sensitivity-based explanations. Users who were only shown the same type across multiple scenarios however did not display any explanation-type related effects compared to those receiving different explanations for the same scenarios. This seems to indicate that users are forming preferences over explanation types on exposure to many of them, rather than especially liking or disliking one when it is the only type they have seen. Furthermore, the dislike of case-based explanations is interesting, as the ‘similarity engine’ description of machine learning could even indicate this might be one of the most faithful explanation types to the functioning of the system. In the lab however, through a speakaloud process undertaken during the task, participants indicated that explaining their similarity to a past case is not, in their minds, treating them sufficiently as an individual.

1.6.2. Debiasing

A major response from the fields of machine learning and statistics to the social impact of algorithmic systems has been to frame the issue in terms of fairness.

Fairness in machine learning has, at the time of writing, been framed in the computer science literature almost entirely in the language of discrimination law. It is through this lens I will outline debates in the literature, aware that this starting point itself might be considered problematic (this will be returned to later). Considering the UK, the focus of discrimination law surrounds membership of a protected group, which in the UK includes (to a varying extent depending on the exact provision examined) age; disability; gender reassignment; marriage and civil partnership; pregnancy and maternity; race;²²³ religion or belief; sex; and sexual orientation.²²⁴

I will firstly look at where some sources of unfairness within this frame of thinking can be found in contemporary machine learning pipelines.

The high demand for labelled data in the context of supervised machine learning—the focus of this thesis—can usually only be met by using data from previous decision-making processes, or from explicitly designed labelling activities. If these data reflect existing, unwanted discrimination in society, for example through racial bias in previous hiring panels, or prejudice in image-labelling crowd-workers, the model that is

²²³ As is usual, race is the term used in legal discussion of discrimination, which does not seek to reify or support any theories that different human races exist. It will be used in accordance with this legal style throughout this thesis with the same intention.

²²⁴ Equality Act 2010, s 4.

1. Hello, World!

learned from it, as in essence a similarity engine, will likely encode these same patterns, risking reproduction of past disparities. Machine learning algorithms are *supposed* to discriminate between data points—they are used to tell instances apart from each other—yet some logics of discrimination, even if predictively valid according to a chosen notion of performance, do not seem societally acceptable.²²⁵

Furthermore, there might be uneven capture of the phenomena being modelled. If some defined sub-groups are historically under-sampled, or exhibit more complicated, nuanced or under-evidenced patterns compared to others, models might exhibit differential performance. It is not practically possible to have data on all individuals, quantifying or classifying all factors important to some social phenomenon. People, or aspects of their lives, are always missing. This skew fast makes its way into data-driven systems.

Data are also often cleaned, categorised and transformed before their use in subjective ways. Feature engineering, where input variables are transformed to make them more amenable to modelling,²²⁶ has crucial downstream impact on the behaviour of machine learning systems. It emphasises aspects of certain variables through augmentation, aggregation and summarisation of characteristics whilst downplaying others. For instance, aggregating those who subscribe to different branches of religious doctrine (eg Catholic, Protestant; Shia, Sunni) within a single overarching religion (Christian, Muslim) might collapse distinctions which are highly relevant to questions of fairness and discrimination within certain contexts. Including a standard deviation of a characteristic as an input variable will make it easier for a machine learning model to emphasise divergence from a constructed average. As with many issues in machine learning, the political nature of this classifying and sorting has long been recognised.²²⁷ Categorisation does not just label people, it can create groups and alter future outcomes,²²⁸ just as feature engineering can in machine learning.²²⁹

Model choice itself can be political. Neural networks or random forests are more amenable to capturing synergy between variables than linear regression. Use of regression methods, for example, might omit important contextual variance or interaction effects key to understanding how a phenomenon manifests on-the-ground, per-

²²⁵ For a discussion of the political philosophy of ‘fair’ machine learning and this acceptability, see Binns, ‘Fairness in Machine Learning: Lessons from Political Philosophy’ (n 131).

²²⁶ Domingos (n 77); see earlier 1.3, p. 35.

²²⁷ See eg Bowker and Star (n 35).

²²⁸ Ian Hacking, ‘The looping effects of human kinds’ in David Premack, Dan Sperber and Ann J Premack (eds), *Causal cognition: A multidisciplinary debate* (Oxford University Press 1995); Bernard E Harcourt, *Against prediction: Profiling, policing, and punishing in an actuarial age* (University of Chicago Press 2006).

²²⁹ Antoinette Rouvroy, ‘Technology, virtuality and utopia’ in Mireille Hildebrandt and Antoinette Rouvroy (eds), *Law, Human Agency and Autonomic Computing* (Routledge 2011).

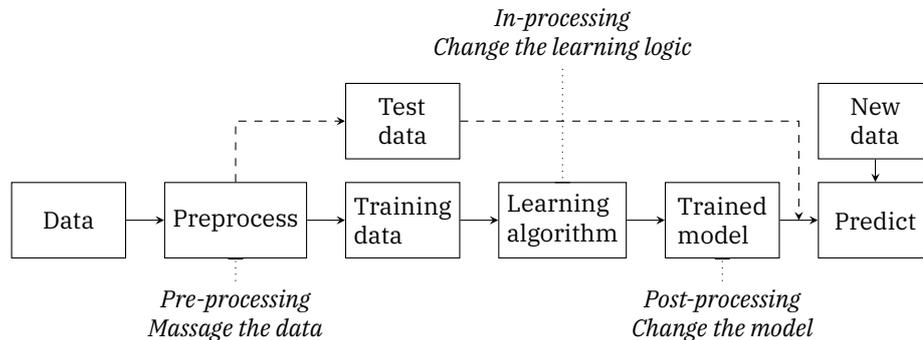


Figure 1.2.: Three approaches to debiasing in the context of the machine learning pipeline (Figure 1.1). Diagram by author.

haps for certain subgroups—as the literature on intersectionality warns.²³⁰ Within a model family, further hyperparameters must be specified. Higher regularisation parameters penalise complexity in a model, which might help it generalise but might trade-off for certain complicated or rare patterns not being retained. Different evaluation mechanisms for models emphasise different aspects of performance.²³¹ Unfortunately, ‘neutral’ choices in machine learning systems do not exist—candidates for these, such as software defaults, are best thought of as arbitrary.

Finally, once built, model deployment may introduce additional fairness issues. The extent to which a model may have different impacts on different groups may only become evident once that model is put into a decision-making system; for instance, the setting of thresholds for positive and negative outcomes could have significant consequences for different groups which may not be evident by merely studying the model itself. The introduction of an algorithmic system may also provide spurious justification for decisions which would otherwise have been more open to challenge under a purely human decision-making process.²³² It might also (similarly to the labelled data problem above) create its own future data, and if this data is problematic, discrimination issues that may have only been mild at the time of model training might be exacerbated over time.²³³

To tackle some of these challenges, computational techniques to prevent machine

²³⁰ Theories around intersectionality highlight that marginalisation based on race and gender are more than the sum of their parts when they occur together. See originally Kimberle Crenshaw, ‘Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics’ [1989] U. Chi. Legal F. 139.

²³¹ Nathalie Japkowicz and Mohak Shah, *Evaluating learning algorithms: A classification perspective* (Cambridge University Press 2011).

²³² Linda J Skitka, Kathleen L Mosier and Mark Burdick, ‘Does automation bias decision-making?’ (1999) 51 *International Journal of Human-Computer Studies* 991 DOI: 10/bg5rb7.

²³³ Ensign, Friedler, Neville, Scheidegger and Venkatasubramanian (n 48).

1. Hello, World!

learning methods from perpetuating some forms of bias have been proposed in recent years by research communities such as discrimination-aware data mining (DADM) and ‘fairness, accountability and transparency in machine learning’ (FATML). Such communities have particularly coalesced at the fringes of computing conferences in workshops such as *FAT/ML*, *FATREC*, *FairUMAP*, *Ethics in NLP*, *FairWare*, *FATWEB*, as well as the fledgling umbrella conference *ACM FAT**.²³⁴

Fairness-aware machine learning methods involve altering usual data science processes in order to correct these forms of bias. They can operate at several stages, including pre-processing, in-processing and post-processing (see Figure 1.2).²³⁵ In each case, the aim is to induce patterns that do not lead to discriminatory decisions despite the possibility of biases in the training data. One recent work states that ‘[f]airness-aware machine learning algorithms seek to provide methods under which the predicted outcome of a classifier operating on data about people is fair or non-discriminatory based on their protected class status’.²³⁶

As already alluded to, anti-discrimination law has particularly motivated these communities, who have attempted to formalise these requirements for mathematical implementation. For instance, heuristics such as the US Equal Employment Opportunity Commission’s ‘80% rule’, which provides a suggested level of permissible disparity between protected groups and the general population, have been used to set parameters for fairness-aware models.²³⁷

There are multiple ways to define fairness formally in machine learning contexts. Most measures focus on differences in treatment between protected and non-protected groups, but there are multiple ways to measure differences in outcomes. These include: *disparate impact* or *statistical/demographic parity*, which requires classification propensities not to differ between individuals who are only distinguished by protected group membership;²³⁸ *accuracy equity*, which constraints the accuracy of

²³⁴ An updated list can be found at <https://fatconference.org/links.html>.

²³⁵ Sara Hajian and Josep Domingo-Ferrer, ‘A Methodology for Direct and Indirect Discrimination Prevention in Data Mining’ (2013) 25(7) *IEEE Trans. Knowl. Data Eng.* 1445 DOI: 10/f4xxs6.

²³⁶ Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton and Derek Roth, ‘A Comparative Study of Fairness-Enhancing Interventions in Machine Learning’ [2018] arXiv preprint (<http://arxiv.org/abs/1802.04422>), 3.

²³⁷ See eg Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger and Suresh Venkatasubramanian, ‘Certifying and removing disparate impact’ in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015) DOI: 10/gfgrbk.

²³⁸ Some of these measures have obvious shortcomings. In particular, disparate impact has been criticised because it fails to account for discrimination which is explainable in terms of legitimate grounds. See further Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold and Richard Zemel, ‘Fairness through awareness’ in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS ’12)* (2012) DOI: 10/fzd3f9. For instance, attempting to enforce equal impact between men and women in recidivism prediction systems, if men have higher reoffending rates, could result in women remaining in prison longer despite being less likely to reoffend.

a predictive model to be the same between groups;²³⁹ its sibling *conditional accuracy equity*, where the accuracy of a predictive model must not differ between individuals in different protected groups conditional on their predicted class;²⁴⁰ *equality of opportunity*, which considers whether each group is equally likely to be predicted a desirable outcome conditional on the observed ‘base rates’ for that group;²⁴¹ and *disparate mistreatment*, a corollary which considers differences in false positive rates between groups.²⁴² Other measures focus not just on actual outcomes and their relation to true/false positives/negatives, but on counterfactual scenarios wherein members of the protected groups are instead members of the non-protected group (ie a woman classified by the system should get the same classification she would have done had she been a man, including all the differences in her life that might have brought).²⁴³ This effectively echoes considerably older calls to more carefully consider the role of causality in digital discrimination.²⁴⁴ Other work focuses on similarity, requiring the user to specify some distance matrix giving a similarity score for data record to every other data record, and endeavouring to treat similar cases similarly.²⁴⁵

Each of these approaches are arguably reasonable ways to measure fairness within a narrow context or problem frame. One might therefore hope that a fair system would or could satisfy all of these constraints. But unfortunately, recent work has formally proven that it is impossible for a model to satisfy several of these constraints at the same time, except in exceptional cases which some argue are unlikely to hold in the real world.²⁴⁶ As a result, choices between the different measures will have to be made. In some cases it may be more important to focus on differences between positive classifications (eg in loan applications), and therefore an ‘equality of opportunity’ measure

²³⁹ Angwin, Larson, Mattu and Kirchner (n 82); William Dietrich, Christina Mendoza and Tim Brennan, *COMPAS risk scales: Demonstrating accuracy equity and predictive parity* (Northpointe 2016).

²⁴⁰ Dietrich, Mendoza and Brennan (n 239).

²⁴¹ Moritz Hardt, Eric Price and Nati Srebro, ‘Equality of Opportunity in Supervised Learning’ in DD Lee, M Sugiyama, UV Luxburg, I Guyon and R Garnett (eds), *Advances in Neural Information Processing Systems 29* (Curran Associates, Inc 2016).

²⁴² Zafar, Valera, Gomez Rodriguez and Gummadi (n 87)

²⁴³ Matt J Kusner, Joshua Loftus, Chris Russell and Ricardo Silva, ‘Counterfactual Fairness’ in I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett (eds), *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc 2017) (<http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf>).

²⁴⁴ Oscar H Gandy, ‘Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems’ (2010) 12(1) *Ethics and Information Technology* 29 DOI: 10/bzwqrx.

²⁴⁵ Dwork, Hardt, Pitassi, Reingold and Zemel (n 238).

²⁴⁶ Chouldechova (n 86); Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns and Aaron Roth, ‘Fairness in Criminal Justice Risk Assessments: The State of the Art’ [2018] *Sociological Methods & Research* DOI: 10/gfgt87; Kleinberg, Mullainathan and Raghavan (n 86). The ‘exceptional cases’ centre on the base rates between protected groups being the same—that is, for example, that Black individuals really do have some statistically different propensity to recidivate than other groups in the population. This is a contentious argument both empirically and otherwise and one which the author believes is worthy of further scrutiny and potential reframing.

1. Hello, World!

might be preferable; in others, the cost of a false negative might be higher (eg the risk a violent criminal might pose to the public). The choice of a particular fairness measure therefore ought to be sensitive to the context.

As a side-note, there are times where fairness-aware machine learning does not utilise fairness characteristics. For example, as described in the algorithmic war story around natural language processing,²⁴⁷ word embedding systems can incorporate undesired stereotypes picked up from the language use in the data sources drawn upon during training. Approaches for ‘fairness’ in this domain have been restricted to identifying particularly heinous stereotypes and attempting to sterilise them by, for example, attempting to collapse a latent dimension thought to represent gender differences.²⁴⁸ Image recognition systems also suffered from issues that are not amenable to imposing fairness constraints, such as those in the *Google Gorillas* case,²⁴⁹ where a particular type of failure mode that occurred with a certain cultural sensitivity. It might not be the case that offence would be taken if Black individuals were disproportionately misclassified as trees and White individuals as cars, for example, and as such any kind of formal or generalised approach to this would have to encode culturally specific notions of misclassification likely to cause harm or outrage. Natural language processing and speech recognition come with their own sets of issues, failing to recognise vernacular or accents shared by certain demographics,²⁵⁰ or struggling in general with ‘under-resourced languages’.²⁵¹

Setting aside definitional problems or domain tensions with the applicability of definitions at all, fairness-aware machine learning techniques are increasingly seen as desirable, viable and even in some cases recommended by regulators or implied by legislative documents. Within European contexts, non-discrimination and data protection are rights enshrined in the EU Charter of Fundamental Rights,²⁵² and both potentially relate to the risks of unfairness inherent in machine learning applications.²⁵³ The recitals of the General Data Protection Regulation 2016 (GDPR)—a law that will be amply discussed in this thesis—refer in particular to fairness-aware data mining tech-

²⁴⁷ Section 1.4.4, p. 45.

²⁴⁸ See Bolukbasi, Chang, Zou, Saligrama and Kalai (n 118).

²⁴⁹ Section 1.4.4, p. 45.

²⁵⁰ See eg Su Lin Blodgett and Brendan O’Connor, ‘Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English’ in *Presented at FAT/ML 2017, Halifax, Canada* (2017) (<https://arxiv.org/abs/1707.00061>).

²⁵¹ See generally Laurent Besacier, Etienne Barnard, Alexey Karpov and Tanja Schultz, ‘Automatic Speech Recognition for Under-Resourced Languages: A Survey’ (2014) 56 *Speech Communication* 85 DOI: 10/gfpgrw.

²⁵² Charter of Fundamental Rights of the European Union [2012] OJ C326/391 (Charter) art 8; Charter, art 21.

²⁵³ Raphaël Gellert, Katja de Vries, Paul de Hert and Serge Gutwirth, ‘A Comparative Analysis of Anti-Discrimination and Data Protection Legislations’ in Bart Custers, Toon Calders, Bart Schermer and Tal Zarsky (eds), *Discrimination and privacy in the information society* (Springer 2013).

nologies and organisational measures.²⁵⁴

These tools may be useful within their narrow contexts, although they do, *prima facie*, miss several core issues. Such debiasing systems do not address issues of cumulative disadvantage or stigmatisation from individual decisions and choices which may interact in difficult ways difficult to reduce to fairness definitions.²⁵⁵ They do not deal with deeper challenges that result from business models who specifically are designed to optimise, including some societal dynamics but externalising some out of scope of their thinking or operations.²⁵⁶ They do not deal with the potential to manipulate, or the power which accumulation of data and models may bring. To some extent, the formalising of fairness in this way can be seen analogously to the history in civil rights scholarship of small academic circles marginalising more radical or transformative voices.²⁵⁷ Insofar as they render a complex, entrenched and enduring social issue a mere technical fix, it is somewhat unsurprising that such topics have become a mainstay of the industry labs at the world's largest technology firms.

1.6.3. Accounting

The A-for-accountability in FATML has considerably less research behind it than the F-for-fairness or the T-for-transparency.

Broadly, we can see the concept of accountability as torn between being a *virtue* or a *mechanism*.²⁵⁸ Particularly in US political science scholarship, 'being account-

²⁵⁴ 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 (General Data Protection Regulation) [2016] OJ L119/1 (GDPR) recital 71, stating that 'In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to [...] prevent, *inter alia*, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect'. This reflects some not inconsequential grammatical changes made by deed in May 2018 by the Council. See Corrigendum to 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 (General Data Protection Regulation) (OJ L 119, 4.5.2016) [2018] OJ L127/2 (GDPR Corrigendum). See generally Michael Veale and Lilian Edwards, 'Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling' (2017) 34(2) *Comput. Law & Secur. Rev.* 398 DOI: 10/gdhrtm.

²⁵⁵ Oscar H Gandy, *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage* (Routledge 2009); Gandy, 'Engaging Rational Discrimination' (n 244); Lina Dencik, Arne Hintz, Joanna Redden and Harry Warne, *Data Scores as Governance: Investigating uses of citizen scoring in public services* (, Data Justice Lab, Cardiff University 2018).

²⁵⁶ Overdorf, Kulynych, Balsa, Troncoso and Gürses (n 132).

²⁵⁷ Richard Delgado, 'The Imperial Scholar Revisited: How to Marginalise Outsider Writing, Ten Years Later' (1992) 140 *U. Penn. L. Rev.* 1349.

²⁵⁸ Mark Bovens, 'Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism' (2010)

1. Hello, World!

able' is a virtue encompassing the behaviour of public (or increasingly, consequential private) agents. When used in this way, it is especially likely to get mixed with fairness, transparency, and other broad, loose and vague political desiderata. In continental European, Canadian, British and Australian scholarship however, it is often considered more narrowly as a social mechanism or relation, whereby, in the words of Mark Bovens, accountability is 'a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences'.²⁵⁹ To avoid confusion with the other terminology, it is proposed here that it is the *latter* concept, accountability as a mechanism, which is most useful in the study of algorithmic systems—if only because accountability is in the computational discourse primarily placed next to other desiderata (as in FATML) rather than treated as a conceptual umbrella.

Computationally, this process has often been seen through computational methods of asking questions of actors and getting mathematically-assured answers in return from them. 'Computer science accountability' has been described as 'a technical concept about making sure that software produces evidence allowing oversight and verification of whether it is operating within agreed-upon rules'.²⁶⁰ Such evidence usually comes in the form of some mathematical proof, which might prove for example that a certain event occurred, a certain piece of information was held, or that a certain system exhibits particular bounds to its behaviour.

Technical tools for this type of accountability can include technologies such as software verification, where a proof is acquired that a software object in practice matches some specification; cryptographic commitments, a file (such as a hash) that can be matched to a digital object to see, for example, if it has subtly changed; or a zero-knowledge proof, which creates a proof that a certain condition was fulfilled without revealing further information, such as what the information that led to the triggering of that condition was.²⁶¹ Primarily, the threat model in this case (the thing that a user would want accountability *for*) would be the threat of a 'good' model having been swapped by a malicious adversary for a 'bad' one. This swapping attack could, for example, be used to swap a system which had been audited for one of the fairness conditions described above²⁶² for a model which had not been (but may perform in

33(5) West European Politics 946 DOI: 10/frq37t.

²⁵⁹ Mark Bovens, 'Analysing and assessing accountability: A conceptual framework' (2007) 13(4) European L.J. 447 DOI: 10/b4hmbf.

²⁶⁰ Deven R Desai and Joshua A Kroll, 'Trust But Verify: A Guide to Algorithms and the Law' (2018) 31 Harv. J.L. & Tech. 1, 10.

²⁶¹ Joshua A Kroll, Joanna Huey, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson and Harlan Yu, 'Accountable Algorithms' (2017) 165 U Pa. L. Rev. 633.

²⁶² Section 1.6.2, 65.

ways more favourable to the controller’s balance sheet). Mechanisms to prevent this cryptographically, linked to debiasing techniques, have been proposed.²⁶³

A separate but related field in computing is less concerned with cryptographic forms of accountability, and more concerned with clear ways to record and analyse information that would be useful for tracing accountability in a system. These include frameworks around ‘model governance’ and ‘decision provenance’, which borrow concepts from semantic web technologies and from data provenance to attempt to track training, usage, decisions and outcomes that are connected to algorithmic systems.²⁶⁴ This can be seen as a form of transparency aiding accountability, as well as an example of a model-centric explanation (as described in the previous section).

It is important to note that conceptions of accountability in the FATML literature largely neglect the agency of those seeking to hold powerful actors to account. Following Hirschman, we can understand two main courses of action: *exit* and *voice*, where individuals would either try to leave the operation of a system, or attempt to control and amend it.²⁶⁵ The notion that with technical mechanisms to highlight issues, such as model swapping, accountability as a whole will follow, is a problematic one which will be revisited in the legal sections of this thesis.

* * *

Explaining, debiasing and accounting for systems have emerged as three central, self-defined pillars concerning computer scientists’ roles around governing machine learning in society. Now the context has been set through both these ‘solutions’ and the motivating war-stories previously outlined, I will outline the research questions underpinning this thesis, and the structure that it will take going forward.

1.7. Research Questions and Method

In light of these concerns around machine learning systems, and the emerging responses by the computing field in particular, I now move to outline the research questions, method and structure of this thesis.

²⁶³ Niki Kilbertus, Adria Gascon, Matt Kusner, Michael Veale, Krishna P Gummadi and Adrian Weller, ‘Blind Justice: Fairness with Encrypted Sensitive Attributes’ in *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)* (2018) (<http://proceedings.mlr.press/v80/kilbertus18a.html>).

²⁶⁴ See eg Jatinder Singh, Jennifer Cobbe and Chris Norval, ‘Decision Provenance: Capturing data flow for accountable systems’ [2018] arXiv preprint (<https://arxiv.org/abs/1804.05741>); Vinay Sridhar, Sri-ram Subramanian, Dulcardo Arteaga, Swaminathan Sundararaman, Drew Roselli and Nisha Talagala, ‘Model Governance: Reducing the Anarchy of Production ML’ in *USENIX ATC’18* (USENIX Association 2018).

²⁶⁵ Albert O Hirschman, *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States* (Harvard University Press 1970).

1. Hello, World!

There are many ways in which ‘machine learning that matters’ might be regulated. Regulation here is understood as a ‘broad set of attempts to control an environment using control systems’.²⁶⁶ Falling within this are different regulatory modes acting as constraints on actors in the environment, which include architectural modes (eg how infrastructure is designed), social norms, economic forces and law, with its ex post sanctions.²⁶⁷ These modes are also interdependent: law can shape infrastructures, economic forces can shape legal enforcement, or social norms can shape preferences and behaviour in markets. They are also not necessarily commensurable: a lack of legal enforcement cannot usually be made up for through enhanced social norms alone. Added to this, some regulation might rely on technological development, such as the FATML approaches outlined above. Problems look for solutions, solutions look for problems, and the landscape fast becomes messy and causally tangled.

This thesis initially focuses on legal aspects, in the sense of formal rules, while being mindful of their context amidst and interwoven with the other modes. These too can take a variety of forms. Parallels here can be drawn to internet regulation, where three broad strands of governance can be discerned.²⁶⁸ Self-regulatory approaches at different levels are underpinned by a belief that the speed and dynamism of digital technologies in particular do not lend themselves to more coercive modes of regulation. Yet given the importance of algorithmic systems in determining societal outcomes, self-regulation can be seen to lack some of the substantive and procedural qualities or safeguards that citizens (and importantly for incumbent governments, voters) might expect, leading to calls for a reassertion of sovereignty in the digital realm and an imposition of powers and mechanisms with a legal quality. Yet given heavy levels of industry capture, particularly in areas often arcane for legislators such as technology law, maintaining alignment with the desires of the general public is difficult. This is both the case in emerging areas (such as, perhaps, the governance of algorithmic systems) where individuals have not yet formed clear policy preferences on the topic, but can equally be said of much older and more pervasive technology regulation issues, such as network neutrality.²⁶⁹ As a consequence, a third model, that of *co-regulation*, has seen a mix of the two previous models, with regulation created with a range of different voices around the table. To some extent, this model also has helped with some of the jurisdiction-spanning issues of the internet, yet suffers from the clear question—

²⁶⁶ Christopher T Marsden, *Network Neutrality: From Policy to Law to Regulation* (Manchester University Press 2017) DOI: 10/cxt8.

²⁶⁷ Lawrence Lessig, *Code and Other Laws of Cyberspace* (Basic Books 1999).

²⁶⁸ Ian Brown and Christopher T Marsden, *Regulating Code* (MIT Press 2013) 2–3.

²⁶⁹ Network neutrality, ‘the latest phase of an eternal argument over control of communications media’, can be traced back explicitly to the late 1800s in relation to laws governing telegraph systems, but the early 1800s or earlier when considering the concept of ‘common carriage’ in transport. See generally Marsden, *Network Neutrality: From Policy to Law to Regulation* (n 266).

who is invited to the table, and whose voices matter?

Self-regulation, state regulation and co-regulation have all been discussed in the context of machine learning systems.²⁷⁰ Yet within Europe, data protection is enshrined as a fundamental right, and as an omnibus framework relating to all processing of personal data, touches heavily upon machine learning. Consequently, there is already a statutory regime with significant history to analyse, and one which has been recently updated with the passing of the GDPR in 2016 and its subsequent enforceability as of May 2018. As an omnibus and quite detailed fundamental right, this is worthy of significant analysis in the context of machine learning technologies. Indeed, to analyse self-regulation or co-regulation *without* an understanding of how data protection law does, should and could apply would likely lead to deeply flawed analysis indeed, whilst the same cannot be said so easily the other way around. As a result, Chapter 2 first asks [RQ1]: **to what extent might the current legal regime—European data protection law in particular—serve to successfully regulate harms from algorithmic systems?**

Chapter 2 can be seen as seeking a useful fit between machine learning, algorithmic harms, and the data protection framework. Whether such a fit is always smooth, clear or consistent is a separate research question, and one considered in chapter 3. In that Chapter, I investigate *whether* the European data protection framework is congruent with the technical characteristics of machine learning. The core question is [RQ2]: **whether practices and technologies surrounding machine learning stress this legal framework, and if so, where, why and how?**

The thesis then pivots to consider a different aspect of machine learning governance: not the ‘bird’s eye’ view of pan-European, omnibus data law, but the narrow and context-specific views of data practitioners who may themselves be trying to undertake their work with varying concern for the law, the aims and values of their organisations, and their own concerns and moral precepts. Given the tools discussed in section 1.6, one might assume that practitioners at this level were becoming increasingly well equipped to undertake this task. Yet a core assumption of these tools is that they will be useable and useful to the practitioners making design and deployment decisions today. Very few of the papers and methods already discussed reference any such background research, nor do they assess tools in real organisational environments or even on live problems. The enforceability of not only the law, but

²⁷⁰ For self-regulation, see eg the ‘tenets’ of the Partnership on AI <https://www.partnershiponai.org/tenets/> (archived: <https://perma.cc/8GP3-2EKD>), the IEEE Ethically Aligned Design initiative and associated IEEE P7000 Standards Projects at <https://ethicsinaction.ieee.org>; Association for Computing Machinery, *ACM Code of Ethics and Professional Conduct* (ACM 2018) (<https://perma.cc/2UY9-U8YK>). For state regulation, see section 2. For co-regulation, see eg Department for Digital, Culture, Media & Sport, *Data Ethics Framework* (n 16).

1. Hello, World!

the self- and co-regulatory methods already outlined does depend at least in part on the capacity of practitioners on the ground to cope with issues which may include accountability and discrimination, and this capacity is currently unknown. As a result, this thesis asks [RQ3]: **how congruent are the assumptions and framings of contemporary computational tools designed to tackle social issues in algorithmic systems with real-world environments and constraints?**

Finally, drawing these questions together brings this thesis to a broader, more normative question [RQ4]: **what practical actions might help society better govern machine learning systems?** This question is both an act of research in and of itself, as well as an act of science–policy synthesis with practitioners in mind.

In sum, the four core questions in this thesis:

1. To what extent might the current legal regime–European data protection law in particular–serve to successfully regulate harms from algorithmic systems? (*RQ1*, Chapter 2, *The Law of Machine Learning?*)
2. Do practices and technologies surrounding machine learning stress this legal framework, and if so, where, why and how? (*RQ2*, chapter 3, *Data Protection’s Lines, Blurred by Machine Learning*)
3. How congruent are the assumptions and framings of contemporary computational tools designed to tackle social issues in algorithmic systems with real-world environments and constraints? (*RQ3*, chapters 4, *Coping with Value(s) in Public Sector Machine Learning*; 5, *Unpacking a tension: ‘Debiasing’, privately*)
4. What practical actions might help society better govern machine learning systems? (*RQ4*, all chapters)

The methods deployed in this thesis vary by question and chapter, but have some overarching commonalities. Firstly, the thesis is both exploratory and normative at its heart. Its exploratory nature is due in part to it being written at a time of agenda-setting around machine learning and governance, where people were (and are still) grappling with framing the problems and reaching for and testing a range of governance tools for efficacy and for fit. It does not enter into a long thread of theory amenable to additional building block research questions, although it does engage in dialogue (and attempt to draw together) many disparate threads from different disciplines that have been engaging in highly relevant lines of inquiry. Its normative nature stems from its topic: the political, value-laden issues which are of social concern. These cannot be analysed in a non-normative manner. For example, both the triggering conditions for any remedies in law and the extent and efficacy of remedies all place burden and

responsibility on different actors depending on how they are configured. The aim of this thesis is to propose functional paths to understand and mitigate some of the challenges of machine learning systems, and the challenges as the author sees them are presented within. It is surely not the only way in which these could be presented or discussed, but where they are, an attempt has been made to provide wider context and contesting viewpoints to give the reader as clear and broad a view possible of the issues in question.

Chapters 2 and 3 adopt typical methodologies from academic legal research. These classic legal desk research methods include an analysis of legal scholarship, legislative documents and preparatory documents, case law both at the EU level and, at times, at the UK level, as well as documents from relevant public bodies, such as regulators. Based on these descriptive documents, evaluative analysis within and between legal instruments is undertaken, as well as analysis between legal instruments and computing research in particular, which is reviewed where relevant in each section. The connections made between computing research and data protection law are most starkly made in chapter 3, which provides several wholly novel analytic and interdisciplinary contributions to the study of data protection law. These analyses are helped by a range of empirical case studies and vignettes to illustrate particular harms or technological deployments that might be of concern. Some of these have already been outlined in section 1.4, and some of these are revisited and analysed further in chapter 2, while chapter 3 utilises case studies of Apple's Siri system and Transport for London's Wi-Fi analysis programme (section 3.1.3), and automated lipreading systems (section 3.3) in order to explore relevant tensions in the law. This evaluative work leads within each relevant section to normative recommendations that either stem from the interactions of law and technology (eg sections 3.2.5, 3.3.5) or which are proposed as remedies to the issues identified (eg sections 2.4, 3.1.5).

Chapter 4 sidesteps from the classic academic legal-analytical methods above to undertake empirical work in the area of machine learning in practice. This section reports on semi-structured interviews with public sector machine learning practitioners undertaken in order to understand how they cope with machine learning issues they are experiencing in systems they are building, deploying or maintaining today. The method for this work, which is outlined further in the relevant chapter (see section 4.5), draws upon interview methods and practices from public administration augmented with those from human-computer interaction. This is necessary as while the former field is used to interviews with 'elites', and the latter used to interviews concerning technology on-the-ground, there has historically not been a clear domain

1. Hello, World!

combining the two.²⁷¹ These interviews are then analysed qualitatively, through an organising principle of ‘public sector values’ from the public administration literature (not to be confused with principles relating to the ethics of algorithms discussed earlier²⁷²), and core challenges for computing researchers concerned with the social implications of machine learning are derived. These interviews are analysed in the context of the computing-related approaches to the social impact of machine learning already described in section 1.6. This leads into the following and final core chapter, chapter 5, which expands on one selected aspect of the unearthed tensions from the previous chapter, analysing how it has been approached to date in the computational literature and normatively proposing means to rectify it.

The thesis concludes with a chapter of synthesis and conclusion (chapter 6), drawing together the four core chapters with an agenda for researchers, practitioners and policy-makers.

A short note on interdisciplinarity This thesis is intended to be a deeply interdisciplinary effort, drawing upon research from multiple fields and attempting to create knowledge that connects and spans them. Interdisciplinarity is a word that is often thrown around without much consideration for its form. The work here is inspired by two particular identifiable ‘modes’ of interdisciplinarity.²⁷³ The first is the *subordination service* mode, where one discipline serves to ‘fill the gaps’ in the other(s).²⁷⁴ Throughout this thesis, computer science research is brought in to supplement the lack of consideration for the formal characteristics of technologies often present in analysis of law and policy. In that sense, computer science ‘serves’ the other domains, and is drawn upon when it is needed. However, the most constructive form of interdisciplinarity this thesis uses is the *agonistic-antagonistic* mode.²⁷⁵ By placing law and policy studies against work in computer science, I hope to demonstrate how the combination can be used fruitfully and with rigour to study the tensions and the limits in each other. The ambition is to demonstrate a particular mode of research and thinking in technology law and policy (and the intersection of ‘science, technology, engineering and public policy’ more generally)—a mode that might even characterise a revitalised ‘interdiscipline’—and to see such a mode become more accepted and widespread in

²⁷¹ It could be argued that science and technology studies fills this gap, which in part it does, however the aims of this piece of work are more in line with human–computer interaction where such interviews are designed to directly inform design needs of future technologies and relevant research directions, rather than reflect more broadly on socio-technical systems and their politics.

²⁷² See section 1.5.

²⁷³ See generally Andrew Barry and Georgina Born, ‘Interdisciplinarity: Reconfigurations of the social and natural sciences’ in *Interdisciplinarity* (Routledge 2013).

²⁷⁴ *ibid* 11.

²⁷⁵ *ibid* 12.

future work.

What qualifies me to undertake this kind of research? One reason interdisciplinarity is hard is that it requires researchers to internalise a considerably wider array of content, theoretical scaffolding and methods than they might do sticking to one field alone. The difficulty of doing this as a rigorous solo pursuit is real. In my case, I have sought to make this research possible through i) collaborative networks and ii) by exposing myself to a wide array of viewpoints through events. The former set of networks are indicated through the publications I have listed as being part of this thesis. The work within this thesis is my own, made possible and improved by working with and by having deep conversations and exchanges with colleagues from across disciplines and around the world (as outlined above, p. 1). This has been supported heavily by the workshops and collaborative events I have had the privilege to attend. As a hot topic, there is plenty of scoping activity happening that I have been involved in around the world. In the annex (p. 301), I have tabulated the 100+ workshop-like events I have attended that have exposed me to research and viewpoints across disciplines and fields. The collective discussions and signposts from these events played a large part in informing the arguments and sources drawn upon in this work.

Part II.

Data Protection Law

2. The Law of Machine Learning?

The crossover of AI and law, within conferences such as the International Conference of AI and Law (ICAIL) and the International Conference on Legal Knowledge and Information Systems (JURIX), as well as in journals such as *Artificial Intelligence and Law*, has predominantly, historically, concerned debates and studies relating to the use of artificial intelligence for advancing legal reasoning, inference, information retrieval or practices.²⁷⁶ While much of this debate was focussed with symbolic AI, statistical methods such as machine learning were deployed, particularly in an attempt to cope with the ‘open-textured’ nature of legal rules.²⁷⁷ While some researchers *have* discussed issues of bias or potential bias in such systems, such as the lack of knowledge from female perspectives in child custody expert systems,²⁷⁸ in general this debate did not move from the possibilities²⁷⁹ or limitations²⁸⁰ of using AI within law to using law to govern AI. In the few places the debate did meander in that direction, largely around medical expert systems, it was generally not seen as a priority to regulate, ‘due to [expert systems’] limited availability and the intervention of physicians in [their use].’²⁸¹

The legal components of this thesis focus on data protection law, but data protection is not the only candidate for a law which touches upon machine learning issues. The (much maligned and derided) Database Directive might serve to provide *sui generis* property rights over datasets key to machine learning, and potentially an argument could be made that machine learning systems themselves might warrant such protection.²⁸² Several pieces of proposed European legislation also touch almost expli-

²⁷⁶ See generally Trevor Bench-Capon and others, ‘A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law’ (2012) 20(3) *Artificial Intelligence and Law* 215 DOI: 10/gc7mhr.

²⁷⁷ See eg Trevor Bench-Capon, ‘Neural Networks and Open Texture’ in *Proceedings of the 4th International Conference on Artificial Intelligence and Law* (ICAIL ’93, ACM 1993) DOI: 10/fw933t; John Zeleznikow, ‘The Split-Up project: Induction, context and knowledge discovery in law’ (2004) 3 *Law, Probability & Risk* 147 DOI: 10/bt4rcx.

²⁷⁸ Lilian Edwards, ‘Modelling Law Using a Feminist Theoretical Perspective’ (1995) 4(1) *Information & Communications Technology Law* 95 DOI: 10/bxwnwf.

²⁷⁹ Richard Susskind, *Expert Systems in Law* (Clarendon Press 1987).

²⁸⁰ Philip Leith, ‘Fundamental Errors in Legal Logic Programming’ (1986) 29(6) *The Computer Journal* 545 DOI: 10/bzh3hq.

²⁸¹ Frank D Nguyen, ‘Regulation of Medical Expert Systems: A Necessary Evil’ (1993) 34 *Santa Clara L. Rev.* 1187, 1188. On the limited uptake of expert systems in medicine, see Yang, Zimmerman, Steinfeld, Carey and Antaki (n 179).

²⁸² Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protec-

2. The Law of Machine Learning?

citly upon machine learning: the fairness in business-to-business platforms regulation (on transparency in certain ranking and recommender systems);²⁸³ and the draft Copyright Directive and the draft Terrorist Content Regulation (on data-driven content takedown).²⁸⁴ Competition law around data-driven power,²⁸⁵ consumer law,²⁸⁶ varying liability regimes,²⁸⁷ public law²⁸⁸ and investigatory powers rules²⁸⁹ are also all highly relevant to certain questions around machine learning technologies in different areas and sectors.

Given the alternatives, why (European) data protection? Other than the (reasonable) argument that one thesis cannot do everything, there are several motivations for this. Firstly, it is wide, effectively spanning all private sector activity and most public sector activity apart from security and policing (which are covered by separate regimes). Secondly, I do not think it is too bold to say that most important machine learning systems involve some processing of personal data, a wide concept that will be discussed throughout this thesis. This might be, for example, because they are trained or queried with data relating to individuals, or the results are applied to individual situations. Such processing of personal data in the EU immediately triggers data protection law; the same wide and varied applicability cannot be said of the regimes above. Thirdly, it is new—or newish—with both significant and long-awaited revisions to the 1995 EU law now enforceable as of May 2018, and new data protection laws constantly appearing throughout the world to try and safeguard rights and freedoms in a digital age.²⁹⁰

tion of databases [1996] OJ L77/20 (Database Directive).

²⁸³ Commission, 'Proposal for a Regulation of the European Parliament and of the Council on promoting fairness and transparency for business users of online intermediation services' COM(2018) 238 final.

²⁸⁴ Commission, 'Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market' COM(2016) 593 final; Commission, 'Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online' COM(2018) 640 final.

²⁸⁵ Ania Thiemann, Pedro Gonzaga and Maurice E Stucke, *DAF/COMP(2016)14: Big Data: Bringing competition policy to the digital era - Background note by the Secretariat* (Paris, 2011) ([https://one.oecd.org/document/DAF/COMP\(2016\)14/en/pdf](https://one.oecd.org/document/DAF/COMP(2016)14/en/pdf)); European Data Protection Supervisor, *Privacy and Competitiveness in the Age of Big Data: The Interplay between Data Protection, Competition Law and Consumer Protection in the Digital Economy* (EDPS 2014); Autorité de la concurrence and Bundeskartellamt, *Big Data and Competition* (2016); Bundeskartellamt, *Big Data und Wettbewerb* (2017).

²⁸⁶ Natali Helberger, Frederik Zuiderveen Borgesius and Agustin Reyna, 'The Perfect Match? A Closer Look at the Relationship between EU Consumer Law and Data Protection Law' (2017) 54(5) *Common Market Law Review* 1427.

²⁸⁷ Eric Tjong Tjin Tai, 'Liability for (Semi)Autonomous Systems: Robots and Algorithms' in Vanessa Mak, Eric Tjong Tjin Tai and Anna Berlee (eds), *Research Handbook on Data Science and Law* (Edward Elgar 2018) DOI: 10/gfsq6x.

²⁸⁸ Marion Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power' (2018) 376(2128) *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* DOI: 10/gdxt27; Jennifer Cobbe, 'Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making' [2018] Available on SSRN DOI: 10/gd25bq.

²⁸⁹ Nóra Ni Loideain, 'A Bridge too Far? The Investigatory Powers Act 2016 and Human Rights Law' in Lillian Edwards (ed), *Law, Policy and the Internet* (Hart 2018).

²⁹⁰ See generally Graham Greenleaf, 'European' Data Privacy Standards Implemented in Laws Outside

Lastly, and perhaps most importantly, a potential use of data protection to govern machine learning, or indeed, a potential clash has been highlighted by a range of commentators.²⁹¹ This research seeks to clarify this and other salient tensions before a misrepresentation of them can damage policy debate and development.

2.1. A Regulatory Mixtape

Data protection is a fundamental right in the European Union, enshrined in the Charter of Fundamental Rights of the EU,²⁹² which itself was granted legally binding effect in the Lisbon Treaty.²⁹³ In the Charter, it appears *separately* from the right to respect for private and family life, commonly called the ‘right to privacy’.²⁹⁴ This might surprise some. Scholars and the media both commonly conflate the two notions, such as calling data protection regulations ‘privacy laws’, or equating ‘privacy protection’ with data protection.²⁹⁵ Indeed, insofar as data protection is a regime designed to prevent unjustified processing of personal details whilst pragmatically allowing its processing where rights and freedoms are safeguarded, it might appear to be a spin-off or even an equivalent to the (separate) right to protection against unjustified interferences in private life.²⁹⁶ The intricate relationship(s) between data protection, privacy, and fundamental rights more broadly are however not the subject of this thesis. Indeed, it has already been a core topic of more than one,²⁹⁷ and these will likely not be the last to consider this issue.

Furthermore, while many fundamental rights formalise, enshrine or publicise long and broadly held notions in the public consciousness, ‘data protection’ seems somehow less intuitive and compelling. The proverbial human on the Clapham omnibus can likely tell us (with varying faithfulness to the case law) what they think their rights

Europe’ (2018) 149 *Privacy Laws & Business International Report* 21 (<https://ssrn.com/abstract=3096314>).

²⁹¹ For an account of a potential high-level clash, see eg Tal Zarsky, ‘Incompatible: The GDPR in the Age of Big Data’ (2017) 47 *Seton Hall L. Rev.* 995. For those looking to use data protection to govern machine learning, see section 2.2, p. 91.

²⁹² Charter, art 8.

²⁹³ Treaty of Lisbon amending the Treaty on European Union and the Treaty establishing the European Community, signed at Lisbon, 13 December 2007 [2007] OJ C306/1 (Lisbon Treaty). Data protection is also a treaty right under Consolidated version of the Treaty on the Functioning of the European Union [2016] OJ C202/1 (TFEU) art 16.

²⁹⁴ Charter, art 7.

²⁹⁵ See Gloria González Fuster, *The Emergence of Personal Data Protection as a Fundamental Right of the EU* (Springer 2014) DOI: 10/gfgrb2 270, who documents many cases of this in practice.

²⁹⁶ Paul De Hert and Serge Gutwirth, ‘Data Protection in the Case Law of Strasbourg and Luxembourg: Constitutionalisation in Action’ in Serge Gutwirth, Yves Poullet, Paul De Hert, Cécile de Terwangne and Sjaak Nouwt (eds), *Reinventing Data Protection?* (Springer 2009) DOI: 10/d4n22h 4.

²⁹⁷ González Fuster (n 295); Orla Lynskey, *The Foundations of EU Data Protection Law* (Oxford University Press 2015).

2. The Law of Machine Learning?

Article 7

Respect for private and family life

Everyone has the right to respect for his or her private and family life, home and communications.

Article 8

Protection of personal data

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.

Figure 2.1.: Articles 7 and 8, Charter of Fundamental Rights

to property or rights not to be tortured or enslaved look like. It seems less likely they would be able to answer the same question about their fundamental right to data protection with similar clarity. So, what is it?

This regime did not come out of nowhere, and it is not just a supranational invention of ‘Brussels’. German Länder (in particular, the state of Hesse), France and Sweden were among the first to develop what we now know as data protection back in the 1970s. The Council of Europe formalised the notion into a document now known as Convention 108,²⁹⁸ which served to guide legislators in many countries across the world on the form that data protection was crystallising into.²⁹⁹ Convention 108 has recently been subject to modernisation, adopting many—but not all—features of European law which will be discussed in this thesis.³⁰⁰

Data protection, whether as a subset of privacy,³⁰¹ an overlapping right with privacy, a right enabling *both* privacy and other fundamental rights, or something different en-

²⁹⁸ Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (opened for signature 28 January 1981, entered into force 1 October 1985) 108 ETS.

²⁹⁹ See generally González Fuster (n 295).

³⁰⁰ Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (opened for signature 10 October 2018) 228 CETS.

³⁰¹ Bear in mind that the the right to respect for private life does not solely concern informational privacy, but covers many other aspects, such as intimacy and physical access to bodies and space.

tirely,³⁰² is distinguished by a clear concern with *control* over personal data, autonomy, and ‘informational self-determination’. Control in data protection can be seen in both negative and positive lights: through a positive ‘proactive manifestation of individual autonomy’, such as consenting or utilising rights such as access, erasure or objection, or a negative sense, where users’ autonomy is protection from subversion over those in relative positions of power.³⁰³

How is this fundamental right operationalised in the EU? Its essence is primarily secured through more specific laws which, in relatively intricate ways, govern the relationship between individuals and those controlling data concerning them, and provide for the positive and negative dimensions of control discussed above through an enforcement mechanism centred on independent regulators. The most well-known of these, both in the EU and globally, is the 2016 General Data Protection Regulation 2016 (GDPR),³⁰⁴ the primary (although not sole) aim of which can be seen as safeguarding the essence of the fundamental right to data protection.³⁰⁵ Accompanying these in Europe are the ePrivacy Directive³⁰⁶ (under politically-charged reform into a new Regulation at the time in writing), the so-called Law Enforcement Directive concerning data processing in criminal contexts,³⁰⁷ and a variety of national implementations of data protection, such as the UK’s Data Protection Act 2018.³⁰⁸ These may incidentally help secure the essence of other rights—privacy, non-discrimination, and so on—but their primary focus should likely be seen as data protection.

The ninety-nine articles of the GDPR—the main pillar of contemporary European data protection law—contain a wide variety of diverse regulatory instruments and provisions, novel definitions and concepts, many of which are characterised by the many shades of grey they exhibit in a changing technological landscape. Although I believe this ‘regulatory mixtape’ resists easy summarisation,³⁰⁹ a foolish attempt will now be

³⁰² For a discussion, see eg Maria Tzanou, ‘Data Protection as a Fundamental Right next to Privacy? ‘Reconstructing’ a Not so New Right’ (2013) 3(2) International Data Privacy Law 88 DOI: 10/gfbzpf.

³⁰³ Jef Ausloos, ‘The Right to Erasure: Safeguard for Informational Self-Determination in a Digital Society?’ (Doctoral dissertation, KU Leuven 2018) 53.

³⁰⁴ GDPR.

³⁰⁵ Ausloos (n 303) 61.

³⁰⁶ Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) [2002] OJ L201/37 (ePrivacy Directive).

³⁰⁷ Directive (EU) 2016/680 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 by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA [2016] OJ L119/89 (Law Enforcement Directive). The UK Government calls this the Law Enforcement Directive, but other authors have called it the Police Directive or the Police Data Protection Directive.

³⁰⁸ Data Protection Act 2018.

³⁰⁹ See De Hert and Gutwirth (n 296) 3, who state that it ‘is impossible to summarise data protection in two

2. The Law of Machine Learning?

made to do so.

European data protection law concerns the concept of ‘personal data’, which is broadly defined as any information relating to an identified or identifiable natural person (known as a *data subject*).³¹⁰ A second category of actor, the *data controller*, is a natural or legal person (or other body) who determines, alone or jointly, the purposes and means of the processing of personal data.³¹¹ *Processing* includes almost anything you can imagine doing with personal data: collecting it, transforming or organising it, storing it, consulting it, erasing it or anonymising it.³¹² Data protection broadly gives (positive control) rights to data subjects and places (negative control) obligations on data controllers. These are designed to uphold overarching data protection principles,³¹³ which in summary state that personal data must be

- processed lawfully, fairly and transparently;
- collected for specific, explicit and legitimate purposes and not be processed for further, incompatible, purposes (purpose limitation);
- adequate, relevant and limited to that necessary for the above purposes (data minimisation);
- accurate;
- kept in a form permitting identification for no longer than necessary (storage limitation)
- processed in a manner ensuring appropriate security

Furthermore, the data controller is accountable for these principles, and must be able to demonstrate compliance with them (the ‘accountability principle’).³¹⁴

All processing requires a *lawful basis* to be established before it can be carried out. One of these is consent—separate from other matters, freely given and withdrawn just as easily³¹⁵—but that sits (equally) among necessity for contact, so-called ‘legitimate

or three lines. Data protection is a catch-all term for a series of ideas with regard to the processing of personal data’.

³¹⁰ GDPR, art 4(1) (‘processing’ means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction’).

³¹¹ GDPR, art 4(7).

³¹² GDPR, art 4(2).

³¹³ GDPR, art 5(1).

³¹⁴ GDPR, art 5(2).

³¹⁵ GDPR, art 7.

interests' of the data controller, vital interests of the data subject (ie to save a life), compliance with a legal obligation, or necessity for a task in the public interest or official authority vested in the controller.³¹⁶ These bases must relate to specific purposes the controller has outlined in advance, and communicate to the data subject alongside other information relevant to the processing.³¹⁷ If certain categories of sensitive data are to be processed, there are additional related hurdles to overcome.³¹⁸ Data cannot broadly be processed for new purposes unless these purposes are 'compatible', which requires consideration of links between purposes, the context and power relationships between subject and controller, the sensitivity of the personal data in question, the possible consequences of this processing and the existence of technical and organisational safeguards.³¹⁹

As alluded to earlier, data subjects can trigger certain rights. For example, they can request, electronically and usually freely, a copy of their data and metadata about its processing or transfer—a subject access request (SAR) as the provision became called,³²⁰ the erasure³²¹ or rectification³²² of certain data, or they can object to or restrict certain processing they do not wish to occur but did not explicitly consent to, such as that which the data controller deems to be within their legitimate interests or their public task.³²³ They also have additional rights relating to decision-making concerning them, which will be outlined further below,³²⁴ as well as rights to be notified in the cases of certain high-risk data breaches.³²⁵

A broad system of enforcement underpins these rights and obligations, relating to one (or more) independent supervisory authorities—data protection authorities (DPAs)—per member state. These authorities—the UK's is the Information Commissioner's Office (ICO)—have several powers, including the ability to fine controllers up to 4% of their global turnover, or EUR 20 million, whichever is higher, to compel information or compel controllers to stop a certain processing activity. The specific nature of the powers differs by DPA. Some, like the ICO, have stronger abilities than others around investigation and the acquisition of information from data controllers. Data

³¹⁶ GDPR, art 6(1).

³¹⁷ GDPR, art 13(1)(c).

³¹⁸ Such sensitive data is 'personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership and [...] genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation'. GDPR, art 9(1).

³¹⁹ GDPR, art 6(4).

³²⁰ GDPR, art 15(3).

³²¹ GDPR, art 17, also known by the political moniker of 'the right to be forgotten'.

³²² GDPR, art 16.

³²³ GDPR, art 18; GDPR, art 21.

³²⁴ See section 2.2.1, p. 2.2.1.

³²⁵ GDPR, art 35.

2. The Law of Machine Learning?

subjects also have the ability to empower a non-profit organisation to take their issue forward for them, or to go directly to a court rather than through a supervisory authority.³²⁶

Added to this are several upstream provisions which are partially or fully self-regulatory in nature. DPAs can authorise certain certification systems relating to data controllers or processors,³²⁷ or codes of conduct relating to sectors,³²⁸ which can be used to demonstrate compliance with the GDPR. The data controller must also integrate the principles above into the technical and organisational fabric of the processing system they are deploying, a provision referred to as ‘data protection by design’,³²⁹ and if processing is determined to be ‘high risk’, then a controller must carry out a data protection impact assessment in advance.³³⁰ If the controller determines themselves that they cannot mitigate the risk, they must consult the data protection authority.³³¹ These have been described as ‘meta-regulatory’ in nature—that is to say they act ‘as a means for the state to make corporations responsible for their own efforts to self-regulate’.³³² Overseeing this, for certain categories of data controllers, must be a specific individual designated as a data protection officer (DPO),³³³ who should ‘directly report to the highest management level of the controller or the processor’.³³⁴

This is a quite staggering array of tools which approach a wide array of issues. It is argued elsewhere that together, these tools attempt to safeguard the essence of the fundamental right to personal data protection, among other fundamental rights.³³⁵ While they might not all be available at any one time—many of them, such as the right to erasure, are heavily qualified—the available rights and obligations should, in order to sufficiently ensure the Charter right to data protection,³³⁶ be configured as such to ‘provide a minimum level of ‘control’ over personal data’ at all times when such control might be threatened by informational power structures.³³⁷

³²⁶ See further section 2.3.2, p. 119.

³²⁷ GDPR, art 42; see further section 2.4.3, p. 134.

³²⁸ GDPR, art 40.

³²⁹ GDPR, art 25.

³³⁰ GDPR, art 35.

³³¹ GDPR, art 36.

³³² Reuben Binns, ‘Data protection impact assessments: A meta-regulatory approach’ (2017) 7(1) International Data Privacy Law 22 DOI: 10/cvct, 29.

³³³ GDPR, art 37.

³³⁴ GDPR, art 38(3).

³³⁵ Ausloos (n 303) 61.

³³⁶ Charter, art 8.

³³⁷ Ausloos (n 303) 61.

2.2. Data Protection Rights and Machine Learning

As seen in the introductory chapter of this thesis, machine learning systems have been thought to reconfigure informational power structures in ways which might cause varied harms to individuals and groups. Given the role of data protection outlined above in promoting control around data-driven systems and in a data-saturated society, many have turned to it as a lens through which machine learning and its related harms might be successfully regulated. Its newfound notoriety can perhaps be most effectively illustrated by its international appeal. The United Nations Secretary-General has recommended ‘updating and applying existing regulation, particularly data protection regulation, to the artificial intelligence domain’.³³⁸ Questions on the applicability of parts of the GDPR to profiling and ‘Big Data’ systems have been discussed in legislatures outside the EU around the world, including the US Congress.³³⁹ A range of policy reports on machine learning from outside the EU have taken the GDPR as a framework to build upon.³⁴⁰ The California Consumer Privacy 2018,³⁴¹ incidentally, looks significantly like the GDPR. The change in the winds has been surprising—for many, particularly in the US, the GDPR was considered an EU ‘innovation-killer’. It now appears, at least in part, to look more appealing by the day.³⁴²

2.2.1. Automated decision prohibitions

Individuals have long been disturbed at the idea that machines might make decisions for them which they could not understand or countermand; a vision of out of control authority which derives from earlier notions of unfathomable bureaucracy found everywhere from Kafka’s *Trial* to Terry Gilliam’s *Brazil*. Such worries have emerged from the quotidian world (credit scoring, job applications, speeding camera tickets) as well as the emergent, fictional worlds of technology (wrongful arrest by *Robocop*, 2001: *A Space Odyssey*’s HAL, automated nuclear weapons launched by accident in *Wargames*). Outside of science-fiction concepts, concerns about effectively automated-decisions have often arisen in popular culture. Recurring bank worker, receptionist and holiday

³³⁸ Secretary-General of the United Nations, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/73/150)* (United Nations 2018).

³³⁹ See eg Facebook, Inc, Letter from Facebook Inc. to Chairman Greg Walden: “House Energy and Commerce Questions for the Record” (June 2018) (<https://perma.cc/K6TM-W2N2>) 122.

³⁴⁰ See eg Dillon Reisman, Jason Schultz, Kate Crawford and Meredith Whittaker, *Algorithmic Impact Assessments* (AI Now Institute 2018) (<https://perma.cc/H79W-JN8F>) 13; David G Robinson and Miranda Bogen, *Automation & the Quantified Society* (Upturn and the Netgain Partnership 2017) (<https://perma.cc/QB98-ARFF>).

³⁴¹ 2018 Cal. Legis. Serv. Ch. 55 (A.B. 375).

³⁴² See eg Mark Scott, ‘How Big Tech Learned to Love Regulation’ (*POLITICO*, 11th November 2018) (<https://perma.cc/CQ5N-J3Q9>).

2. The Law of Machine Learning?

rep character *Carol Beer* in the British comedy show *Little Britain* popularised the now common catchphrase ‘computer says no’ in relation to her tendency to deny individuals even the most trivial requests on the basis of machine output.

Automated decision-making systems may have surfaced as a public and scholarly concern only recently, but countries across Europe have had omnibus provisions on automated decision-making for many years longer than might be expected. Some law explicitly describing automated decisions have been in force since 1978, notably in France. A response from Commission nationale de l’informatique et des libertés (CNIL), now France’s data protection authority, stated in 1987 to the International Conference of Data Protection and Privacy Commissioners (ICDPPC), resonates so clearly in the context of the algorithmic war-stories above³⁴³ that it is worth reproducing:

An increasingly prevalent trend is to appreciate the value of a person from the study of his behavior. [...] For example, a consumer credit company may want to verify, before granting a loan to an individual, that they have not given rise in the past to a cash flow incident or any recovery procedure.

Similarly, what will be the attitude of a banker facing a client asking for a mortgage, when the bank finds that the interested party has already had a dispute with another credit institution?

What will be the attitude of a business leader against a candidate for employment, when this employer notices the person concerned had a dispute with a previous employer, even though the courts would have sided with the employee? All the more so in the case of a conflict which has revealed the trade union views of those concerned.

Another example: in a given municipality, housing organisations might be tempted to memorise disputes and to create local files of bad payers or bad tenants. People who have had some difficulties of payment could be banned definitively from accessing housing in their area.

[CNIL] always ensures that no decision is taken on the sole basis of a profile. [...]

In research on artificial intelligence, the idea has emerged of developing expert systems that would resolve disputes brought before a judge. These expert systems are programs that replicate the reasoning of a human expert whose knowledge and experience has been coded and stored in a knowledge base. If there is a problem, the computer system selects the set of

³⁴³ Section 1.4.

rules that might apply to this problem. [...] Can the application of artificial intelligence be transposed into the field of justice? [...]

French law has laid down a fundamental principle which seems to us must be maintained, whatever the progress of technology; a principle which guarantees the citizen against the risk of dehumanised justice.³⁴⁴

The provisions referred by CNIL in French law³⁴⁵ made their way into European data protection law through the initial moves to harmonise provisions across the continent in what was to become the Data Protection Directive 1995 (DPD). While the expansion of these provisions were not aimed at machine learning systems, they were aimed at the general sophistication of models of knowledge in decision-support processes, and in particular expert systems, the debates around which, as the above CNIL quote emphasises, in some ways heavily echo those being had around machine learning today. The European Commission noted in 1992 during the Directive's drafting process that 'the result produced by the machine, using more and more sophisticated software, and even expert systems, has an apparently objective and incontrovertible character to which a human decision-maker may attach too much weight, thus abdicating his own responsibilities.'³⁴⁶ As a consequence, the DPD provided that no decision with *legal or similarly significant* effect could be based *solely* on automated data processing without appropriate lawful basis.³⁴⁷ Some EU Member States interpreted this as a strict prohibition, others as giving citizens a right to challenge such a decision and ask for a 'human in the loop' in order to express their point of view.³⁴⁸

The GDPR has moved and polished this old provision into what is now known as Article 22 (*Automated individual decision-making, including profiling*). The provision is as follows:

1. The data subject shall have the right not to be subject to a decision

³⁴⁴ Communication de M. Alain Simon à la conférence annuelle des commissaires à la protection des données (Québec, septembre 1987), reported in Commission nationale de l'informatique et des libertés (CNIL), *8e Rapport au président de la République et au Parlement, 1987* (La Documentation Française 1988) (<https://perma.cc/2NCW-R5Q3>) 243–248 [author translation].

³⁴⁵ Loi no. 78-17 du 6. janvier 1978 relative à l'informatique, aux fichiers et aux libertés, s 3 of which states 'Any person shall be entitled to know and to dispute the data and logic used in automatic processing, the results of which are asserted against him'. See Lee A Bygrave, 'Automated Profiling: Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling' (2001) 17(1) *Comput. Law & Secur. Rev.* 17 DOI: 10.1016/S0267-3649(01)00104-2.

³⁴⁶ Commission, 'Amended proposal for a Council Directive on the protection of individuals with regard to the processing of personal data and on the free movement of such data' COM(92) 422 final–SYN 297, 26.

³⁴⁷ Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [1995] OJ L281/31 (DPD) ¶ 15

³⁴⁸ Interestingly, this provision has been interpreted by some to imply that European systems are more interested in the human dignity of data subjects than the US system: see Jones (n 129).

2. The Law of Machine Learning?

based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:
 - a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
 - b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
 - c) is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

The *prima facie* remedy to harms related to automated decision-making provided by Article 22 is to require a lawful basis before such a decision can be undertaken. This must be either explicit consent,³⁴⁹ necessity for entering or performing a contract,³⁵⁰ or a basis in Member State law.³⁵¹ Where decisions are based on 'special category' data in Article 9, contract is ruled out as a lawful basis (as it is for processing that data at all).³⁵² Without such a basis, the automated decision cannot be taken. This has been read differently in the past in the DPD, with some Member States claiming

³⁴⁹ References to explicit consent are considered to be stronger than references to consent in the GDPR. Explicit consent is also referenced as a means to lift the restriction on the processing of 'special category' data under GDPR, art 9 and to authorise the transfer of data to a third country in the absence of safeguards under GDPR, art 49. The Article 29 Data Protection Working Party (A29WP) have stated in their guidance on consent that explicit consent centres around the provision of a 'statement', which might include a written and signed statement, 'filling in an electronic form, by sending an email, by uploading a scanned document carrying the signature of the data subject, or by using an electronic signature.' A subtle checkbox or a bundled agreement with other terms both seem unlikely to meet the threshold under this definition. See Article 29 Data Protection Working Party, *Guidelines on Consent under Regulation 2016/679 (wp259rev.01)* (2018).

³⁵⁰ The extent of 'necessity' is likely to be a pivotal area for the Court to decide upon in future.

³⁵¹ GDPR, art 22(2).

³⁵² GDPR, art 22(4). cf GDPR, art 6 where contract is possible as a lawful basis, compared to GDPR, art 9 where it is not.

that a lawful basis would remove this right to object, whilst other Member States claim that such lawful basis is necessary to lift a prohibition.³⁵³ The A29WP have claimed in guidance on the matter that Article 22 should be read as a prohibition.³⁵⁴ Yet either way, the remedy provided is at its heart a means to allow data subjects to obtain an alternative decision-making process, and limit the cases where data subjects are not consulted before to actions deemed necessary for contract or by law.

There are secondary remedies present in Article 22, which are represented by ‘safeguards’ when one of the lawful bases for a decision described above *has* been obtained. For cases where Member State law is *not* used to legitimise decisions, data subjects may request ‘at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision’.³⁵⁵ Where Member State law is the basis for lifting the prohibition, safeguards are to be provided for in national implementing legislation.³⁵⁶

There are also tertiary remedies which connect Article 22 to other Articles in the GDPR. The information rights in Articles 13–15 provide that data subjects are told of the ‘existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.’³⁵⁷ Article 22–style decisions are also a trigger for a data protection impact assessment (DPIA) to be undertaken by the data controller—something which will be discussed later.³⁵⁸

None of these remedies however can be understood without analysing the scope and nature of the Article 22 triggering conditions.

³⁵³ Bygrave, ‘Automated Profiling: Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling’ (n 345).

³⁵⁴ Veale and Edwards, ‘Clarity, Surprises, and Further Questions’ (n 254). There remains a distinction however which is illustrated in the difference in wording and national implementation of the mirrored provision in Law Enforcement Directive, art 11, which mentions prohibition explicitly (‘Member States shall provide for a decision based solely on automated processing, including profiling [...] to be prohibited’)—language different from that in the GDPR. Similarly, the UK, in its transposition of the Law Enforcement Directive, clarifies in DPA 2018, s 49 that ‘[a] controller may not take a significant decision based solely on automated processing unless that decision is required or authorised by law.’ Given that Article 29 guidance does not have the force of law, this debate may yet be prised open.

³⁵⁵ GDPR, art 22(3).

³⁵⁶ See eg DPA 2018, s 14. On implementations in other member states, see Gianclaudio Malgieri, ‘Right to Explanation and Algorithm Legibility in the EU Member States Legislations’ [2018] Presented at CPDP 2019, preprint available on SSRN (<https://papers.ssrn.com/abstract=3233611>).

³⁵⁷ GDPR, art 13(2)(f).

³⁵⁸ GDPR, art 35. See further Article 29 Data Protection Working Party, *ADM Guidelines* (n 2) 29.

2. The Law of Machine Learning?

2.2.1.1. Applicability

Article 22 is considerably restricted in its scope, which is governed by certain characteristics of the application rather than by the technologies at play behind the scenes.

Human involvement Article 22 applicability is predicated in part on the extent of human involvement in a decision, and potentially its subsequent execution. There are cases where decisions are both made and executed in a fully automated way. In the tax system for example, an automated mailshot sent to those likely to not pay income tax on time may trigger Article 22.³⁵⁹ In the digital economy, many businesses, such as social media or labour platforms, primarily operate automatically to enable them to scale to userbases of millions. Yet there are many other cases where the intervention or treatment usually by definition involves a human—social work, medical treatment, or promotion in a physical workplace are all mediated by people. These individuals might be taking a decision-support system’s guidance as input, or they might be acting upon its guidance. Either way, there is some end to the automation and some beginning to the ‘human’, and, as noted by several relevant policy bodies,³⁶⁰ where and how this line is drawn is important to understanding Article 22.

A significant concern that has been present in the human factors literature relates to unwarranted over-reliance on machines: one form of ‘automation bias’.³⁶¹ At what point should over-reliance rend a decision ‘solely’ based on automated processing, and thus with the potential to trigger Article 22, given an overriding likelihood of any human in-the-loop to ‘rubber-stamp’ a decision regardless of its qualities.

For a long time, this went without legal or regulatory clarification. Recently however, it was examined by the A29WP, whose guidance sets the scene for an interesting conceptual challenge. The A29WP provides two interesting statements in this regard.

³⁵⁹ Michael Hallsworth, John A List, Robert D Metcalfe and Ivo Vlaev, ‘The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance’ (2017) 148 *Journal of Public Economics* 14 DOI: 10/f96mbc.

³⁶⁰ The Dutch Scientific Council for Government Policy in early 2016 specifically recommended that more attention be paid to ‘semi-automated decision-making’ in the GDPR, in relation to profiling. See *Wetenschappelijke Raad voor het Regeringsbeleid* (n 191) 142. The ICO similarly highlighted this issue in a 2017 consultation, asking ‘Do you consider that “solely” in Article 22(1) excludes any human involvement whatsoever, or only actions by a human that influence or affect the outcome? What mechanisms do you have for human involvement and at what stage of the process?’. See Information Commissioner’s Office, *Feedback Request—Profiling and Automated Decision-Making [v 1.0, 2017/04/06]* (ICO 2017) 20.

³⁶¹ Skitka, Mosier and Burdick (n 232); Jaap J Dijkstra, ‘User agreement with incorrect expert system advice’ (1999) 18(6) *Behaviour & Information Technology* 399 DOI: 10/fsnqm9; Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce and Hall P Beck, ‘The role of trust in automation reliance’ (2003) 58(6) *International Journal of Human-Computer Studies* 697 DOI: 10/cgvddv. The other form is an under-reliance on machines when they are more trustworthy than human decision-makers, for whatever reason.

Firstly, they note that ‘if someone routinely applies automatically generated profiles to individuals without any influence on the result, this would still be a decision based solely on automated processing’.³⁶² This implies that when considering if ‘solely’ applies to an automated system, DPAs should consider how often the system operators disagree with the system outputs and changes or otherwise augments them. It also applies, in line with the principle of accountability, that this should at least be evidenced, if not recorded in detail.

This approach has interesting consequences which the A29WP do not seem to have contemplated, but which I believe are important going forward. If a machine learning system is *claimed* to outperform humans and treated as such—as many of the PR efforts around ‘AI’ currently argue machine learning often does—any human involvement in the process designed to avoid the application of Article 22 this should *necessarily* be expected to be effectively nominal. Unless a system can demonstrate (also within its marketing, documentation and practices) that there is a role for humans, then the system should be regarded as ‘solely’ automated, and where the significance criterion is also met, will require a more ‘human’ decision-system to exist in parallel. (A counterargument to this would be to state that the humans and the machines both make and detect different errors from each other. Yet either way, this would require the developers and vendors of these technologies to be significantly clearer about the limitations of their tools than they currently are in order to hope not to trigger this provision.)

The second relevant statement of the A29WP notes that ‘meaningful human input’ is required rather than a ‘token gesture’ for the system to avoid categorisation as ‘solely’ automated. This second perspective focusses on ensuring the human has, in the words of the A29WP, the ‘authority and competence’ to change the decision. This forms an interesting challenge. It has been noted that where humans are involved in decision-making, they are often in ‘moral crumple zones’, socially and culturally responsible for the errors of complex systems even where, upon careful analysis, blame is much harder to assign.³⁶³ As and if machine systems become better at given tasks, we can expect maintaining non-token ‘authority and competence’ to be a significant social and organisational challenge, further reducing the scope of avoiding Article 22 obligations.

In sum, the A29WP may have unwittingly set up increasingly performant (or well-branded) machine learning technologies as fuel for expanding the scope of Article 22, as well as have created a new avenue for the GDPR to govern organisational and so-

³⁶² Article 29 Data Protection Working Party, *ADM Guidelines* (n 2) 21.

³⁶³ Madeleine Clare Elish, ‘Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction’ (2019) 5 *Engaging Science, Technology, and Society* 40 DOI: 10/gf2t99.

2. The Law of Machine Learning?

cial structures and practices. It is, however, worth emphasising that the A29WP guidance is not binding. While, in practice, the Court of Justice of the European Union (CJEU) does appear to follow their guidance closely (and may do so further with the new guidance of the European Data Protection Board (EDPB), which has now replaced the A29WP),³⁶⁴ we still await Court action to clarify these many blurry points.

Further to, and somewhat tempering, these two statements by the A29WP, there are compelling statistical grounds to practically require a human-in-the-loop when attempting to predict *rare events*. A model that can predict complex rare events is a machine learner's dream, and it is often at the heart of 'AI' products' promises. Yet a common problem with a statistical basis appears when predicting rare events: the heavy number of false positives that are inevitably generated when one tries to do so.³⁶⁵ This is exacerbated by the changing nature of many phenomena, where the issue trying to be modelled is not stationary for long enough to get adequate modelling information about the rare cases in order to predict them in the future.³⁶⁶ The utility of the metaphor of a phenomenon as statistical distribution breaks when such a 'distribution' can only be sampled from a small number of times before it is changed. These false positives tend to outnumber the resources that can be devoted towards acting upon these predictions, and as a result not all can be followed up upon. This is a common challenge in areas such as terrorist detection systems, where some have argued in certain predictive tasks, it is credible to think false positives outnumber true positives by a factor of 10,000.³⁶⁷ A human-in-the-loop to sift and prioritise given a 'longlist' is a typical function in areas like child welfare or policing.

Significance Article 22 also applies only to decisions that have 'legal' or 'similarly significant' effects on an individual. Some would argue this could only apply to systems which make important, binding decisions on things like criminal justice, risk assessment, credit scoring, education applications or employment. The CJEU has, similarly to what counts as *solely* based on automated processing, never been asked to determine the scope of what counts as a *significant* effect—nor are there any cases on this issue pending before the Court at the time of writing.

³⁶⁴ Nadezhda Purtova, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10(1) Law, Innovation and Technology 40 DOI: 10/gd4rmh.

³⁶⁵ T Ryan Hoens, Robi Polikar and Nitesh V Chawla, 'Learning from streaming data with concept drift and imbalance: An overview' (2012) 1(1) Progress in Artificial Intelligence 89 DOI: 10/tx4rz5.

³⁶⁶ See generally J Gama, Indre Žliobaitė, A Bifet, M Pechenizkiy and A Bouchachia, 'A survey on concept drift adaptation' (2013) 1(1) ACM Comput. Surv. DOI: 10/gd893p; Joaquin Quiñero-Candela, Masashi Sugiyama, Anton Schwaighofer and Neil D Lawrence, *Dataset shift in machine learning* (The MIT Press 2009).

³⁶⁷ Timme Bisgaard Munk, '100,000 False Positives for Every Real Terrorist: Why Anti-Terror Algorithms Don't Work' (2017) 22(9) First Monday DOI: 10/cvzf.

This test of significance is challenging in practice for several reasons.

Firstly, automated decisions may well have a different effect on different individuals, and not affect all data subjects uniformly. Some data subjects may be in a situation of precarity or vulnerability that makes them susceptible in ways others are not. Considering the *Sweeney Search* war-story,³⁶⁸ while the results seem intuitively unacceptable, it seems farfetched to suggest that they damaged Professor Sweeney herself. As a professor at Harvard University, she has considerable comparative social and (likely) economic capital, and the successful papers and profile she gained for herself from this search incident mean that this particular incident could even have been seen to personally benefit rather than harm her. The same cannot be said of a (hypothetical) individual living in poverty, applying for a job only to be rejected when an employer, upon searching her name online, was put off by the impugned criminality they saw in response to their query. The significance trigger is contextual, and that context can vary between data subjects for a multitude of reasons depending on the domain.

If one-in-a-hundred individuals who are delivered a highly personalised advert for gambling are significantly affected by it for some definition of significant, yet that vulnerability is difficult or impossible to observe in advance of such a decision being made, would this oblige Article 22 obligations to trigger in advance for all data subjects? What about one-in-a-thousand? Even assuming such propensities are known explicitly is quite a stretch, and asks a lot of the analytic capacity in organisations taking such decisions. As we have seen online, unexpected outcomes or effects of digitally delivered measures are often arguably significant for *someone*: meaning that either Article 22 provisions be extended across sectors and applications widely, or that they be reserved for clearly 'significant' issues, rendering the *ex ante* provisions useless for particularly vulnerable data subjects.

Coping with this contextual variance raises further tricky and unresolved issues, because data controllers have obligations under the GDPR that *require knowledge in advance of whether a decision triggers its provisions*. These include a right to be informed that a decision is being undertaken,³⁶⁹ the provision of information under Articles 13–14 where data is collected but no decision has been made,³⁷⁰ and, where explicit consent is relied upon as a ground to render such a decision lawful,³⁷¹ the need to collect such consent in advance. Terminology around other automated decision-related provisions in the GDPR, including information rights and DPIAs, also uses future-facing language. Information rights require information to be provided on 'envisaged con-

³⁶⁸ Section 1.4.2, p. 42.

³⁶⁹ Wachter, Mittelstadt and Floridi (n 201); Article 29 Data Protection Working Party, *ADM Guidelines* (n 2).

³⁷⁰ See section 2.2.2.2.2, p. 112.

³⁷¹ GDPR, art 22(2)(c).

2. The Law of Machine Learning?

sequences’,³⁷² DPIAs, which explicitly include Article 22–style profiling,³⁷³ talk of processing ‘*likely* to result in a high risk to the rights and freedoms of natural persons’ [emphasis added].³⁷⁴

This future-looking is integrated into the GDPR without fanfare, but has challenging implications. In complex data-driven systems, understanding what the ‘likely’ downstream impacts of upstream representations are is a challenging task. The ‘normal accidents’ hypothesis argues that when systems are complex from many interlinked internal subsystems, and tightly coupled with little lag between one process influencing another, the range of potentially disastrous interactions are impossible to predict and difficult to locate and diagnose when they occur.³⁷⁵ These are precisely the types of data-driven systems being built today, which use data collected with a variety of otherwise unrelated means to shape individuals’ experiences, which in turn influences future data collection. Data subjects are represented, quantified and categorised in different ways throughout this pipeline and with little clear causal connection to outcomes. Indeed and in general, machine learning has been evaluated in terms of ‘what works’ rather than why it works, with little scrutiny being placed on these processes unless they are negatively impacting upon performance metrics.³⁷⁶

A second stressor of the notion of significance is the idea of *cumulative effect*. Some decisions may seem trivial as a one-off event, but are significant in aggregate, potentially resulting in ‘cumulative disadvantage’.³⁷⁷ For example, the biases in natural language processing systems discussed in the Bias in the Toolkit cases³⁷⁸ retrench and add to assumptions and stereotypes widely believed to propagate unfairness. In the case of political advertising and concerns of ‘filter bubbles’,³⁷⁹ it is the cumulative effect, rather than the one-off instance, that is of interest. Some scholars have argued that advertising can never be ‘significant’ in the way Article 22 demands,³⁸⁰ but since their work the A29WP have argued that there may be cases where it might be.³⁸¹ Where individuals’ environments, or even their preferences, are deliberately shaped over time, would *this* be ‘significant’? As a ‘decision’ or ‘measure’ is an an-

³⁷² GDPR, art 13(2)(f); GDPR, art 14(2)(g); GDPR, art 15(1)(h).

³⁷³ GDPR, art 35(3)(a).

³⁷⁴ GDPR, art 35(1).

³⁷⁵ Charles Perrow, *Normal accidents: Living with high risk technologies* (Basic Books 1984).

³⁷⁶ Leo Breiman and others, ‘Statistical modeling: The two cultures’ (2001) 16(3) *Statistical Science* 199 DOI: 10/bd86gq; Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books 2018).

³⁷⁷ Gandy, *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage* (n 255).

³⁷⁸ Section 1.4.4, p. 45.

³⁷⁹ See section 1.4.5, p. 47.

³⁸⁰ Isak Mendoza and Lee A Bygrave, ‘The Right Not to Be Subject to Automated Decisions Based on Profiling’ in Tatiana-Eleni Synodinou, Philippe Jougoux, Christiana Markou and Thalia Prastitou (eds), *EU Internet Law* (Springer 2017) DOI: 10/gfscwg.

³⁸¹ Veale and Edwards, ‘Clarity, Surprises, and Further Questions’ (n 254).

choring point for triggering Article 22, this is unclear. Over time, machine learning systems may affect processes of identity formation, and the cumulative effects of predictive technologies thought to ‘nudge’ users is unclear, and brings difficult to anticipate consequences.³⁸²

This last issue of accumulated effect is a good starting point for the next challenge around Article 22: the focus of its provisions. What, indeed, *is* a ‘decision’?

2.2.1.2. The nature of ‘decisions’

Terry Gilliam’s classic film *Brazil* opens in a clinical, semi-mechanised bureaucracy. A rogue fly is swatted into a printer of citizen records—a literal bug in the machine—resulting in a classification error. Instead of an arrest warrant being produced for the wanted renegade air conditioning mechanic Archibald Tuttle (played by Robert De Niro), innocent cobbler Archibald *Buttle* was bundled away by secret police, incarcerated and fatally interrogated. Buttle’s neighbour, Jill Layton, is similarly impugned with criminality as she reports and investigates Buttle’s wrongful arrest, leading to the chaotic, tangled (and visually stunning) pursuits of the film.

Assuming the world of *Brazil* falls under the GDPR, or a similar regime with an Article 22–inspired provision, at what point was the consequential decision made about Jill?³⁸³ As with all tricky sociotechnical questions, there is no one clear answer. On one hand, decisions were taken each time law enforcement used her profile (a criminal) and applied it in chasing her through the city. They looked at their machines and yes—still, Jill was their target. Yet the source of the automated failure can be traced much further upstream to the automated ‘decision’ to apply that classification to her in the first place. Indeed, in the mechanised bureaucracy of *Brazil*, the consequences of that classification were clear. If such a classification were made manually by an individual rather than by a machine, it would likely be in full awareness of the downstream consequences. In many non-fictional domains, individuals are indeed triaged before further processes are applied to them. They are classified in ways which changes the logic of treatment (potentially algorithmic) in the future. Intelligence services may classify individuals as high risk to security, using this as a proportionality argument to justify more extensive invasion of privacy. Medical services triage individuals as more at risk, and provide differentiated treatment as a result.

Child protection services serve to exhibit real examples of multi-stage algorithmic

³⁸² Karen Yeung, ‘“Hypernudge”: Big Data as a Mode of Regulation by Design’ (2017) 20(1) *Information, Communication & Society* 118 DOI: 10/gddv9j calls this ‘hypernudging’. See generally Hildebrandt, *Smart technologies and the End(s) of Law* (n 68).

³⁸³ These issues might fall under the Law Enforcement Directive instead of the GDPR, but set that aside for now.

2. The Law of Machine Learning?

systems. A stylised example: given limited resources, a child protection agency uses automatically generated risk scores to proposed to rank and prioritise cases to be dealt with most urgently. These cases are sorted into low, medium and high risk bins, which might change over time (for example, as a case sits for a longer time in a lower risk bin). Social workers pick from these bins according to guidelines relevant to their role and expertise. They have some discretion in doing so. Once chosen, additional algorithmic systems are present to help recommend specific courses of action.

In the field of algorithmic bias, this has been described as the importance of considering ‘representational’ harms, usually as the root of ‘allocative’ harms.³⁸⁴ Representational harms are based on how individuals are presented, classified, recorded in models and datasets. Intrinsicly, these can be damaging, reinforcing stereotypes across eg racial lines³⁸⁵ or influencing perceptions based on who is counted and who counts, for example in the case of national measures of violence against women.³⁸⁶ They can also, as in the child protection example above, more directly affect downstream ‘allocative’ action: who gets what, whether it be resources or a particular treatment or course of action.

How we define a ‘decision’ in data protection law strongly influences whether the provisions in Article 22 are effective or internally coherent. In the guidance on this issue, the A29WP do not tackle the issue of what conceptually a decision is, focussing primarily on the broader issue of when Article 22 triggers.³⁸⁷ I argue there are two distinct choices to make here, terming them the *subsidiarity approach* and the *counterfactual approach*.³⁸⁸

Subsidiarity approach In political science, subsidiarity is a neologism coined in the 1930s: a notion that issues should be dealt with at the lowest level that best ensures their resolution. At its core (initially religious in nature), it is about ‘larger associations aiding but not superseding the smaller ones’.³⁸⁹ It has been most popularised through the principle of subsidiarity introduced to the EU in the Treaty of Maastricht, whereby ‘the Union shall act only if and in so far as the objectives of the proposed action cannot

³⁸⁴ Crawford (n 141).

³⁸⁵ Noble (n 140).

³⁸⁶ Merry (n 35).

³⁸⁷ See Article 29 Data Protection Working Party, *ADM Guidelines* (n 2) 8.

³⁸⁸ There are parallels here, although not identical ones, to the concepts of ‘factual causation’ versus ‘scope of liability’ or ‘proximate cause’ in tort law. The notion of separating the two is largely attributed to Leon Green, *The Rationale of Proximate Cause* (Vernon Law Book Company 1927), and the distinction between the two is critically discussed in David Hamer, ‘Factual Causation’ and ‘Scope of Liability’: What’s the Difference?’ (2014) 77(2) *The Modern Law Review* 155 DOI: 10/gfv2qt. Broader possible connections between the notion of a ‘decision’ and tort law are left to future work.

³⁸⁹ Gerald L Neuman, ‘Subsidiarity’ in *The Oxford Handbook of International Human Rights Law* (Oxford University Press 2013) DOI: 10/gfrjh9.

be sufficiently achieved by the Member States, either at central level or at regional and local level, but can rather, by reason of the scale or effects of the proposed action, be better achieved at Union level'.³⁹⁰

Under a subsidiarity approach, a 'decision' could be seen in terms of the nearest point to the moment of allocation where such allocation decision could have been effectively changed—such as the triggering of an investigation, payment of welfare support, or delivery of information such as an advert. It is the last person, or system, who 'presses the button', and who could have realistically pressed it otherwise. In a decision tree, it is the arbiter of the last fork where a different outcome would be possible.

Counterfactual approach A counterfactual approach would be much broader. It would look at *all* points in the decision system where a different choice would have (significantly) changed the expected outcome, usually including the proximity approach point, but typically incorporating more upstream decision points still.³⁹¹ Political scientists commonly analyse entire political systems in terms of the agents who could have changed (or refused) a certain course of action,³⁹² or the points at which those decisions could have been vetoed.³⁹³ For example, a risk score might change the propensity for downstream auditors to investigate a potential case for tax fraud. Following the subsidiarity approach, it is the point where the risk score is used by auditors to initiate an investigation which would be considered the decision. Following the counterfactual approach however there would be additional eligible decision points including, importantly, the point at which the risk score was generated.

The counterfactual approach is important because it gives greater focus on the systems as a whole, and the impacts of their structures on data subjects. The importance of taking a systemic view to algorithms has already been emphasised in this thesis.³⁹⁴ Decisions that are the most proximate to the individual are often unlikely to be the most compelling 'cause' of a harm, which may stem from much further upstream, in the practices of data collection, cleaning, labelling and augmentation, classification and profiling. A counterfactual approach draws attention to the points of system

³⁹⁰ Art 5(3) TFEU.

³⁹¹ This is not to be confused with the term counterfactual, which is used with quite different definitions in areas such as algorithmic transparency (Wachter, Mittelstadt and Russell (n 205)) and algorithmic discrimination (Kusner, Loftus, Russell and Silva (n 243)).

³⁹² See e.g. the 'veto players' approach in George Tsebelis, 'Decision making in political systems: Veto players in presidentialism, parliamentarism, multicameralism and multipartyism' (1995) 25(3) *British Journal of Political Science* 289 DOI: 10/cc622q.

³⁹³ Ellen Immergut, 'Institutions, veto points, and policy results: A comparative analysis of health care' (1990) 10(4) *Journal of Public Policy* 391 DOI: 10/bvr5nc.

³⁹⁴ See section 1.1, p. 27.

2. The Law of Machine Learning?

design and process when choices impact upon hundreds, thousands or even millions of individuals, yet the proximity approach limits its scope to the individual in question, at the point in time where they are materially affected. To this extent, the counterfactual approach shares some similarities with notions of the importance of information in cybernetics. The notion of information as a *difference which makes a difference*³⁹⁵—that some deviation from a state triggers a later action, can be compared to some distinguishing profile, risk score, inference causing downstream downstream.

Furthermore, the counterfactual approach aims at harms before they happen in a damaging or irreversible way. For example, if an individual is stopped-and-searched on the street on the basis of automated decision-making, or if a transaction was blocked at a pivotal moment, or content automatically censored during the middle of an important event, how can that decision be effectively challenged or undone?³⁹⁶ If issues were flagged further upstream, it may however prevent such issues from coming to pass, and give individuals a real opportunity to exercise the remedies provided in a fast-moving context. As it stands, this is often challenging.

Interestingly, upstream decisions through such a counterfactual approach might create a different type of interaction with the ‘significance’ criteria already described earlier. The further upstream it is possible to go in a triaging or risk scoring process, the more cumulative disadvantage³⁹⁷ that would potentially contribute to ‘significance’ would or might be captured. Considering the *Cambridge Analytica* war-story³⁹⁸ through the subsidiarity lens might result in a decision being identified on an ad-by-ad basis, in which case, it might not be reasonable to suggest that any single advertisement would easily swing an individual’s opinion in an election. Considering a decision as whether to apply a profile to an individual or not within a system, leading to several adverts delivered to them over a longer period of time, the likelihood of significance looks larger.

Which approach has the most backing in the GDPR?

The first place to look is the text itself. The text does provide two examples of what such a decision might look like: ‘automatic refusal of an online credit application’ or ‘e-recruiting practices without any human intervention.’ These are highly canonical ‘decisions’ discussed heavily in studies hypothesising solutions to algorithmically propagated discrimination.³⁹⁹ Both are already subject to quite extensive sec-

³⁹⁵ Gregory Bateson, *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology* (Chandler Pub Co 1972).

³⁹⁶ This was also a point raised in the UK Parliament during the debates around the Data Protection Act 2018. See HC Deb 13th March 2018, vol 637, 54.

³⁹⁷ Gandy, *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage* (n 255).

³⁹⁸ Section 1.4.5, p. 47.

³⁹⁹ See section 1.6.2, p. 65.

toral regulation. Additionally, they typify decisions that have traditionally been taken by humans, and which are either increasingly perceived as or feared to be taken by machines.⁴⁰⁰ While, as discussed earlier, these are the sorts of decisions that Article 22 of the GDPR or Article 15 of the DPD has been envisaged to cover for decades,⁴⁰¹ they only reflect a subset of the influences *processes* machine learning is involved in, many of which reflect new or significantly changed tasks (eg advertising, content take-down, customer tracking, information retrieval) that do not always have a clear, previous human-mediated analogue.

Beyond those two examples, there is little more explicit discussion of what a decision consists of beyond that it ‘may include a measure’.⁴⁰² The term measure is not used elsewhere in the GDPR to refer to a data subject-facing activity: it is usually used in reference to ‘technical and organisational measures’, which are seen as safeguards in processing activities, or a measure taken by a Member State, such as a regulation.⁴⁰³ A measure in the sense recital 71 implies would, looking at the Oxford English Dictionary, likely be either a ‘plan or course of action intended to attain some object; a suitable action’ or a ‘[t]reatment (of a certain kind) meted out to a person.’⁴⁰⁴ The former seems more likely, particularly given the nature of the German and French translations ‘Maßnahme’ and ‘mesure’ respectively. The notion of planning seems intuitively more upstream than that of a decision, and can (albeit only weakly) be seen to support a wider variety of decision points, such as those in a counterfactual approach.

Beyond this, in some circumstances, defining a decision though a subsidiarity approach might not ensure what the CJEU established as a principle in *Google Spain*⁴⁰⁵ and *Wirtschaftsakademie*:⁴⁰⁶ ‘effective and complete protection of data subjects’. If defining a decision in subsidiarity terms precludes societally important notions of what is a ‘significant’ effect, particularly through gradual impacts in the digital realm, this seems inherently limiting. Furthermore, this principle has recently been proposed to be expanded more broadly in a parallel direction by Advocate General (AG) Bobek in his opinion on the upcoming CJEU *Fashion ID* case, as ensuring the law provides ‘ef-

⁴⁰⁰ On automated hiring systems, see Miranda Bogen and Aaron Rieke, *Help Wanted—An Exploration of Hiring Algorithms, Equity and Bias* (, Upturn 2018).

⁴⁰¹ See the CNIL statement in section 2.2.1, p. 91.

⁴⁰² GDPR, recital 71.

⁴⁰³ It is however used in a data subject in the context of ‘marketing measures’ in the questions referred to the CJEU in Case C210/16 *Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v Wirtschaftsakademie Schleswig-Holstein GmbH* ECLI:EU:C:2018:388, ¶ 78. The Court did not grapple with this term in its judgement, however.

⁴⁰⁴ ‘measure’, in *Oxford English Dictionary* (2nd edn, Oxford University Press 2014) (<http://www.oed.com/view/Entry/115506>).

⁴⁰⁵ Case C-131/12 *Google Spain v Agencia Española de Protección de Datos (AEPD) and González* ECLI:EU:C:2014:317, ¶ 34.

⁴⁰⁶ *Wirtschaftsakademie* (n 403) ¶ 28.

2. The Law of Machine Learning?

efficient and timely' protection to data subjects.⁴⁰⁷ Given that 'decisions' in the digital realm can have little or no lag time between being taken and their effects (such as adverts or automatic takedown requests), a subsidiarity approach effectively precludes the ability to contest a decision provided for by Article 22. By the time such contestation can occur, the damage has been done. One solution to this is to allow for contestation further upstream, in a context-sensitive way: another argument in favour of a counterfactual approach as opposed to a subsidiarity one.

However, this in turn faces a different challenge: blurring the distinction between *profiling*, *processing* and *automated decision-making* in data protection law. Processing is the broadest term, as previously discussed, encompassing more or less anything that can be done with personal data. Profiling is a subset of processing undertaken 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'.⁴⁰⁸ Automated decisions, as discussed, appear to be a distinct matter, although must be based on 'automated processing, including profiling', implying that while profiling is important, it is not necessary. It is unclear, for example, whether automated decisions evaluating situational or environmental as opposed to 'personal' aspects would fall under Article 22. Either way, decisions and processing are distinct activities. Yet to adopt a counterfactual approach to defining decisions might mean that processing activities themselves can qualify as decisions, creating significant blurring that the CJEU may wish to avoid, if it is ever posed the question.

Either way, defining decisions in this context is a problem for data protection law. A narrow definition appears to arguably provide inadequate protection; a wide definition risks attempting to turn all processing into decision-making, requiring frequent analysis of when a particular informational action would go on to change the world—arguably often impossible in a complex sociotechnical system. It may be the case that a sector-by-sector or application-by-application approach is warranted in this situation in order to locate a proportionate midpoint between these two difficult outcomes.

2.2.2. Information and Explanation Rights

The desire for machine learning systems to explain themselves has become a popular proposed remedy to algorithmic harms in recent years. Earlier in this work, this hype

⁴⁰⁷ Opinion of AG Bobek C-40/17 *Fashion ID GmbH & Co KG v Verbraucherzentrale NRW eV* (joined parties: *Facebook Ireland Limited, Landesbeauftragte für Datenschutz und Informationsfreiheit Nordrhein-Westfalen*) ECLI:EU:C:2018:1039, ¶ 132.

⁴⁰⁸ GDPR, art 4(4).

and the technical approaches designed to respond to it has been described.⁴⁰⁹ In this section, the *legal* grounds for explanation rights will be explored, as well as a broader consideration of its limitations as a remedy.

Individual explanation rights of a sort are commonly found in the *public* sphere in democracies, in the form of freedom of information (FOI) rights against public and governmental institutions. Transparency is seen as one of the bastions of democracy, liberal government, accountability and restraint on arbitrary or self-interested exercise of power. As former US Supreme Court Justice Louis Brandeis (widely credited as discovering or deducing a right to privacy in the US constitution) noted, ‘sunlight is said to be the best of disinfectants; electric light the most efficient policeman.’⁴¹⁰ Transparency rights against public bodies enable an informed public debate, generate trust in and legitimacy for the government, as well as allow individual voters to vote with more information. These are perhaps primarily societal benefits, but citizens can clearly also benefit individually from getting explanations from public bodies via FOI: opposing bad planning or tender decisions, seeking information on why hospitals or schools were badly run leading to harm to oneself or one’s child, and requiring details about public funding priorities are all obvious examples. Indeed, in relation to algorithmic transparency, it is easy to see how FOI rights can be used to request details of models—and indeed, this method has already formed the basis of some academic inquiries in this area.⁴¹¹

2.2.2.1. Access rights

FOI regimes were not designed as a policy response to computer-aided or automated decision-making. A parallel transparency right to FOI with a personal rather than a societal angle, and one of the earliest routes to taming pre-machine learning automated processing, appeared in Europe with the creation of subject access requests (SARs). The GDPR’s subject access right is as follows:⁴¹²

1. The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:

⁴⁰⁹ See section 1.6.1, p/ 54.

⁴¹⁰ Louis Brandeis, *Other People’s Money, and How Bankers Use it* (National Home Library Foundation 1933) 62.

⁴¹¹ See eg Marion Oswald and Jamie Grace, ‘Intelligence, policing and the use of algorithmic analysis: A freedom of information-based study’ (2016) 1(1) *Journal of Information Rights, Policy and Practice* DOI: 10.21039/irpandp.v1i1.16, in a policing context; Robert Brauneis and Ellen P Goodman, ‘Algorithmic Transparency for the Smart City’ (2018) 20 *Yale J. Law. & Tech.* 103 DOI: 10/cncv, in a smart city context.

⁴¹² GDPR, art 15.

2. The Law of Machine Learning?

- a) the purposes of the processing;
 - b) the categories of personal data concerned;
 - c) the recipients or categories of recipient to whom the personal data have been or will be disclosed, in particular recipients in third countries or international organisations;
 - d) where possible, the envisaged period for which the personal data will be stored, or, if not possible, the criteria used to determine that period;
 - e) the existence of the right to request from the controller rectification or erasure of personal data or restriction of processing of personal data concerning the data subject or to object to such processing;
 - f) the right to lodge a complaint with a supervisory authority;
 - g) where the personal data are not collected from the data subject, any available information as to their source;
 - h) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.
2. Where personal data are transferred to a third country or to an international organisation, the data subject shall have the right to be informed of the appropriate safeguards pursuant to Article 46 relating to the transfer.
 3. The controller shall provide a copy of the personal data undergoing processing. For any further copies requested by the data subject, the controller may charge a reasonable fee based on administrative costs. Where the data subject makes the request by electronic means, and unless otherwise requested by the data subject, the information shall be provided in a commonly used electronic form.
 4. The right to obtain a copy referred to in paragraph 3 shall not adversely affect the rights and freedoms of others.

Effectively, as implemented by the GDPR, access rights provide access to both a *copy*

*of personal data undergoing processing and metadata surrounding it.*⁴¹³ The metadata include purposes, recipients, source, storage limit and other factors. The copy of the data covers a wide array of information and seems deliberately left comparatively underspecified in the text in order to facilitate a broad and ‘technology neutral’ implementation.⁴¹⁴

Access rights have been a constant integral component of data protection law, visible in the 1973 principles of the Council of Europe on the matter, and integrated largely (although not universally) into early national versions of instruments that flowed from these discussions.⁴¹⁵ Across Europe, such a right has been present in the DPD 1995⁴¹⁶ and has been migrated and extended within the GDPR into that printed above. Although the US lacks an omnibus notion of data protection laws, particularly in relation to the public sector more broadly, comparable rights did emerge there too in relation to credit scoring in the Fair Credit Reporting Act 1970,⁴¹⁷ and more recently in the California Consumer Privacy Act 2018.⁴¹⁸

The purpose of a right to access, at least in the form it has taken in Europe, has been seen by scholars as multi-faceted.⁴¹⁹ Firstly, access rights have a *sine qua non* flavour to them—accessing data is an indispensable prerequisite to rights such as rectification,⁴²⁰ erasure,⁴²¹ objection⁴²² or the restriction of processing.⁴²³ Without knowing the data that a controller holds about you, effectively making use of those rights seems impossible.⁴²⁴ They also have an enforcement function. Data protection authorities, like many regulators in the digital world, are outgunned, understaffed, and asked to tackle a much broader and at times more nebulous set of issues than they ever have been before.⁴²⁵ Access rights reveal—even by their incompleteness or non-provision—

⁴¹³ Quite confusingly, the CJEU has called the personal data ‘basic data’ and metadata data which ‘relates to the processing of the basic data’. This is quite unwieldy terminology with unclear provenance, so this thesis deviates from the Court’s language. See Case C-553/07 *College van burgemeester en wethouders van Rotterdam v MEE Rijkeboer* ECLI:EU:C:2009:293, ¶¶ 41–43.

⁴¹⁴ Evidence supporting that can be found in the CJEU’s arguments in Nowak. See Case C-434/16 *Peter Nowak v Data Protection Commissioner* ECLI:EU:C:2017:994, ¶ 34.

⁴¹⁵ Jef Ausloos and Pierre Dewitte, ‘Shattering one-way mirrors—data subject access rights in practice’ (2018) 8(1) *International Data Privacy Law* 4 DOI: 10/cwcf, 5.

⁴¹⁶ DPD, ¶ 12.

⁴¹⁷ 15 U.S.C. §1681, *et seq.* See generally Danielle Keats Citron and Frank Pasquale, ‘The scored society: Due process for automated predictions’ (2014) 89(1) *Washington Law Review* 1.

⁴¹⁸ 2018 Cal. Legis. Serv. Ch. 55 (A.B. 375) §1798.100(a).

⁴¹⁹ Ausloos and Dewitte (n 415) 7–8.

⁴²⁰ GDPR, art 16.

⁴²¹ GDPR, art 17.

⁴²² GDPR, art 21.

⁴²³ GDPR, art 18.

⁴²⁴ The importance of the right of access as a prerequisite has been confirmed by the CJEU. See *Rijkeboer* (n 413) ¶¶ 51–52.

⁴²⁵ Christopher T Marsden, ‘Prosumer Law and Network Platform Regulation: The Long View towards Creating OffData’ (2018) 2(2) *G’town L. Tech. Rev.* 376.

2. The Law of Machine Learning?

something about the competence or legality of a controller's processing which can help flag issues to raise with regulators or in the courts. This decentralised form of policing, where individuals themselves are given (mild) tools to compel information provision from even powerful actors, can have important consequences, and has been examined as a form of governance in many sectors.⁴²⁶ The most extreme outcome of this kind so far in data protection is arguably how then-law student Max Schrems' access request against social media giant Facebook led to a series of complaints and legal actions, resulting in the striking down of the EU-US Safe Harbor agreement legitimising international data transfers by the CJEU.⁴²⁷

Access rights are an important transparency provision for machine learning from multiple angles.

Firstly, they allow individuals to access information on the purposes for which their data is being processed. If their data is being used to query or build a machine learning model, this should be described here (and indeed, should generally be also provided without asking). Individuals can then, once aware of this, choose to use complementary rights to attempt to control unwanted uses of their data, such as the rights to *object*, to *restrict processing* or to *erase*.

Secondly, they enable users to access a copy of their personal data being used to query a machine learning model—for example, to test their financial means. This may result in them discovering errors or inaccuracies, which they can use their *right to rectification* in order to correct.⁴²⁸

Thirdly, there are specific parts of the right of access that concerns information to be provided in at least the aforementioned cases of *automated decision-making* already discussed in the preceding section.⁴²⁹ It is these which I now turn to examine.

2.2.2.2. Automated decision transparency

As already discussed, both the DPD and the GDPR contain provisions concerning automated decision-making. At the core of these provisions, now Article 22 of the GDPR, is a remedy that qualified automated decisions are prohibited without a lawful basis, and require safeguards when they do have such a basis. This has already been out-

⁴²⁶ See generally Archon Fung, Mary Graham and David Weil, *Full Disclosure: The Perils and Promise of Transparency* (Cambridge University Press 2007).

⁴²⁷ Case C-362/14 *Maximilian Schrems v Data Protection Commissioner* ECLI:EU:C:2015:650.

⁴²⁸ It should be noted that it is unclear how much power an individual has to 'rectify' highly subjective data points. While an opinion about an individual has been held as personal data, the Court has struggled with this question of how to set conditions on applicability for the right to rectification that make sense in the past. On the question of whether rectification rights apply to exam script responses (they do not), *Nowak* (n 414).

⁴²⁹ See section 2.2.1, p. 91.

lined and discussed in the previous section.⁴³⁰ Yet sitting on top of this prohibition, both linked to Article 22 and also, separately and differently, within Article 15 (and 13–14), is a separate remedy which has brought with it myth, intrigue and intense debate. This is a transparency remedy, which in varying forms has become known as a *right to (an) explanation*.

The issues the designers of this provision seemingly had in mind, such as credit scoring, public or rented housing applications and employment applications, appeared to entail the intuition that challenging a decision, and possibly seeking redress, involves this preceding right to an explanation of how the decision was reached. Focussing on the GDPR, such a right can be found in two flavours.

2.2.2.2.1. Article 22 The first flavour of the right can be found as a *safeguard* to qualifying Article 22 automated decisions with a valid lawful basis. When such decisions are allowed, there is a list of safeguards that should be present. Unfortunately, it was quickly identified after the passing of the GDPR that this list appears in the final legislative text in two different forms.⁴³¹ Recital 71 of the GDPR notes in relation to automated decision-making that [emphasis added]

such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, *to obtain an explanation of the decision reached after such assessment* and to challenge the decision.

In Article 22(3), however, a similar list appears, in which such an explanation right is conspicuously absent.

the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

To further confuse the situation, where such a decision involves processing special category data, Article 22 specifies no list of safeguards at all, making it unclear where these should be drawn from.

It is important to note that in European legislation, the *articles* in the main text are binding on member states but are accompanied by *recitals*, which are designed

⁴³⁰ See section 2.2.1, p. 91.

⁴³¹ Wachter, Mittelstadt and Floridi (n 201).

2. The Law of Machine Learning?

to help states interpret the articles and understand their purpose. Recitals are usually regarded as helpful rather than binding, but this is contested and differs among states.⁴³² Unfortunately—and this is not the only time this occurs in the GDPR—some key matters in the recitals are not echoed in the main text.⁴³³ This is highly problematic in European law, and not what recitals are intended to do. Many political issues in the GDPR were kicked into the long grass of the recitals—presumably for interpretation by the CJEU—when they could not be resolved in the inter-institutional, closed-door ‘trialogue’ negotiations between the European Parliament, Council of Ministers and the European Commission.⁴³⁴

2.2.2.2. Articles 13–15 The second location for an explanation right is found within Articles 13–15 of the GDPR. As described above, Article 15 is a right which is activated upon the request of a data subject, and maximally gives them both a copy of their personal data undergoing processing, and metadata surrounding this processing. Articles 13–14 provide for metadata only, but do not require (nor envisage a facility for) active subject participation in their triggering. Instead, Article 13 concerns what ‘metadata’ (author’s term) should be provided when data is collected directly from a data subject, and Article 14 where data is passed from one controller to another. Article 11 is an important article laying out the conditions and the framework for such communications, such as the appropriate language and timelines for delivery. Article 13–14 obligations carried out in practice are likely most familiar to the average reader in the form of documents such as privacy policies, or in information sheets for participants in research studies.

Article 13–15 all contain the following quite dense clause in the lists of information that must be provided to data subjects:⁴³⁵

the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

⁴³² Tadas Klimas and Jurate Vaiciukaite, ‘The Law of Recitals in European Community Legislation’ (2008) 15(1) *ILSA Journal of International and Comparative Law*, 92.

⁴³³ See for example Veale and Edwards, ‘Clarity, Surprises, and Further Questions’ (n 254) 403, describing the contrast between automated decision-making concerning children, where GDPR, recital 71 presents a clear ban, whilst GDPR, art 22 is silent on the issue. This has caused challenges for regulators trying to interpret the law consistently.

⁴³⁴ Similar relegations to the recitals happened in the domain of data protection impact assessments GDPR, art 35. See further Binns, ‘Data protection impact assessments: A meta-regulatory approach’ (n 332). For a closer look at the legislative process around the explanation right’s omission from the main text, see Wachter, Mittelstadt and Floridi (n 201).

⁴³⁵ GDPR, art 13(2)(f); GDPR, art 14(2)(g); GDPR, art 15(1)(h).

At first glance, this provision would appear to assuage concerns around the recital–article clash which plagued the Article 22–based remedy. Both *meaningful information about the logic involved* as well as *the significance and the envisaged consequences of such processing* appear more substantively broad and rich in comparison to a mere ‘explanation’. It is tagged onto information or access provisions which already get reasonably heavy use, unlike Article 22 which has barely been tested. Furthermore, the use of ‘at least in those cases’ is a hook upon which these rights might be expanded, particularly in light of the obligation on controllers to consider the ‘nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons’ in the measures they take to demonstrate compliance with the regulation.⁴³⁶ Yet it could also be seen to be a redundant phrase, and permit controllers to release information in other situations too—something they were never restricted in doing anyway.⁴³⁷

There are two particular dimensions of ambiguity which scholars have highlighted around the use of these rights to provide an ‘explanation’ of a decision made about an individual.

The first surrounds the timing of this provision, and how that impacts the information that is needed to be provided. Article 13–14 information must be available at the time of data collection (in the case of Article 13) or when data is obtained indirectly (Article 14). Assuming that an automated decision is made on the basis of such information, this means that the provisions for *meaningful information* are necessarily already triggered before a decision has been made. This stands in contrast with the Article 15 access remedy, which, despite sharing exactly the same phrasing, is generally triggered by a data subject after the fact. Usually, this ‘fact’ is data collection, and the answers to the metadata questions in Article 15, such as purpose, length of storage, and recipients, do not change. Yet in the situation where an automated decision concerning a data subject has since been made, should such *meaningful information* contain more subject-specific information, tailored for that individual, or should it only contain the information that was already able to be provided before such a decision has been made (or a system at least queried with a data subject’s specific data)?

Some scholars have argued, primarily with reference to the phrasing in the text, that the ‘envisaged consequences’ phrasing in the text has a forward-looking character, necessarily situating the *meaningful information* as that possible to know prior to a specific decision being made.⁴³⁸ Other scholars have criticised this reading, ar-

⁴³⁶ GDPR, art 24(a).

⁴³⁷ Gianclaudio Malgieri and Giovanni Comandé, ‘Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation’ (2017) 7(4) *International Data Privacy Law* 243 DOI: 10/gddkmf, 250–251.

⁴³⁸ Wachter, Mittelstadt and Floridi (n 201).

2. The Law of Machine Learning?

guing that in reference to the overarching principles of data protection (ie to ensure *fair processing*), the definition of *meaningful* will have a context-specific nature, at times justifying specific information to be provided.⁴³⁹

In absence of relevant jurisprudence—the CJEU have never been referred an issue around either of these articles, or their analogues in the DPD—the most official guidance comes in the form of a document from the A29WP, the former group that provided guidance on data protection law (now replaced with a formal body of the Union, the EDPB).⁴⁴⁰ A29WP Guidelines do not have the formal force of law, as a group designed initially to harmonise application of the law across Member States⁴⁴¹ and ‘on its own initiative, [to] make recommendations on all matters relating to the protection of persons with regard to the processing of personal data in the Community’.⁴⁴² The EDPB’s role is more elaborate than the A29WP’s, but remains rooted in consistency. Yet in relation to the tensions within the law, some of which have been outlined above and some of which will be identified below, the Board’s role cannot be simply seen through the lens of consistency, but at times it is specifying underspecified aspects of the law in absence of a CJEU ruling on the matter. Part of this can be seen in the GDPR, where the tasks of the board include issuing guidelines, recommendations and good practices not only ‘on its own initiative’ in any area of the Regulation,⁴⁴³ but also on complex aspects of specific areas, such as profiling,⁴⁴⁴ erasure⁴⁴⁵ and data breaches.⁴⁴⁶

Implicitly and without fanfare, the A29WP appear to align themselves with the view that the Articles 13–15 right to ‘meaningful information about the logic involved’ provide a ‘more general form of oversight’, rather than ‘a right to an explanation of *a particular* decision’ [emphasis original]. The information should, they argue, consist of ‘simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision, without necessarily always attempting a complex explanation of the algorithms used or disclosure of the full algorithm’. Interestingly, earlier on the A29WP also explicitly note that Article 15, which is triggered by a data subject explicitly seeking information, implicitly after processing has commenced, does not provide the data subject with more information than should have been provided under Articles 13 and 14: that the same phrasing is indicative of the same information, re-

⁴³⁹ Andrew D Selbst and Julia Powles, ‘Meaningful information and the right to explanation’ (2017) 7(4) *International Data Privacy Law* 233 DOI: 10/gddxmz.

⁴⁴⁰ Article 29 Data Protection Working Party, *ADM Guidelines* (n 2).

⁴⁴¹ DPD, 30(1)(a).

⁴⁴² DPD, 30(3).

⁴⁴³ GDPR, art 70(1)(e).

⁴⁴⁴ GDPR, art 70(1)(f).

⁴⁴⁵ GDPR, art 70(1)(d).

⁴⁴⁶ GDPR, art 70(1)(g).

regardless of the timing of the right being different.⁴⁴⁷ Put together, this approach seems designed to fatally damage the chances of generating a personalised *ex post* ‘right to an explanation’ from Article 15(h)—and indeed the example given of ‘meaningful information’ restricts itself to regurgitating back (i) input information provided by the data subject, (ii) relevant information provided by others (eg credit history) and (iii) relevant public information used in the decision (eg public records of fraud).⁴⁴⁸ In other words, it seems suggested that no information about the ‘innards’ of the decision-making process—anything of a decompositional nature⁴⁴⁹—need be given. This intuitively seems *too* restrictive, given that this would be a redundant and overlapping right, providing only data that the data subject was already permitted to access under the access right to a copy of data processed relating to them under Article 15.

There is likely little use in arguing over whether the A29WP is ‘right’ or not, as to do so amounts to guessing what the Court will say. Even given the limited jurisprudence concerning data protection as a whole, it has already not been unheard of for the CJEU to directly contradict the A29WP. Likely the most famous incident is the *Google Spain* case,⁴⁵⁰ where the CJEU overruled the A29WP on aspects concerning the creation of the *right to be delisted*,⁴⁵¹ one of the pre-cursors to today’s *right to erasure*.⁴⁵²

It is also worth noting that the profiling guidance preceded the majority of public concerns around Cambridge Analytica and microtargeting.⁴⁵³ In the case of the UK, and with the support of the ICO, these concerns have led to investigations and reports,⁴⁵⁴ as well as the incorporation of algorithmic explanations as a priority area in the regulator’s technology strategy.⁴⁵⁵ Furthermore, in a UK context, it is unclear whether the A29WP opinions will bear heavily at all on the interpretation of law were the UK to sever ties with the European Union, as this was something left answered in the debates on the then-Data Protection Bill,⁴⁵⁶ and which will likely only be clarified as, when and if UK courts diverge from the CJEU.

* * *

⁴⁴⁷ Article 29 Data Protection Working Party, *ADM Guidelines* (n 2) 15.

⁴⁴⁸ *ibid* 14.

⁴⁴⁹ See above section 1.6.1, p. 54.

⁴⁵⁰ *Google Spain* (n 405).

⁴⁵¹ *Purtova* (n 364) 59–60.

⁴⁵² See generally Ausloos (n 303).

⁴⁵³ See the corresponding algorithmic war-story in section 1.4.5, page 47.

⁴⁵⁴ Information Commissioner’s Office, *Democracy Disrupted? Personal Information and Political Influence* (ICO 2018) (<https://perma.cc/2M2N-QQSX>); Information Commissioner’s Office, *Investigation into the use of data analytics in political campaigns* (ICO 2018) (<https://perma.cc/2X2U-X6Q4>).

⁴⁵⁵ Information Commissioner’s Office, *Technology Strategy, 2018-2021* (ICO 2018) (<https://perma.cc/7RJ5-DAB6>).

⁴⁵⁶ See eg HL Deb 13th November 2017, vol 785, 1863; HL Deb 13th December 2017, vol 787, 1576.

2. The Law of Machine Learning?

Beyond a broad overview, and instead of speculating on the details of implementation as the Court might see it, it is more useful to critically examine some of the characteristics of these remedies (in their various potential forms) in relation to their efficacy in addressing the harms outlined earlier in the thesis. In many cases, I now argue, we find them lacking.

2.3. Problems of data protection individualism

I now turn to some higher level problems with the aspects of data protection that might remediate machine learning harms or challenges described above.

2.3.1. The transparency fallacy

The utility of individual, rights-based data subject explanations has never really been justified in terms of practical efficacy in relation to the broad range of algorithmic decisions individuals appear to be faced with. In many cases it is easy to suspect that what the data subject wants is *not* an explanation—but rather for the disclosure, decision or action simply not to have occurred.

Indeed, considering the few modern EU legal cases we have on controlling algorithmic governance, an explanation has not usually been the remedy sought. An interesting example is the seminal CJEU *Google Spain*⁴⁵⁷ case mentioned earlier, which introduced the ‘right to be delisted’⁴⁵⁸ (which since has partially morphed into the right to erasure in the GDPR⁴⁵⁹). This case still remains one of the few instances of individual algorithmic harm to have come to the highest EU court. In this case, the claimant, Mr Costeja González, asked Google to remove a link to an old and outdated page in a newspaper archive recording his long-repaid public debt that formed the top returned link for searches on his name. Mr Costeja González’s (successful) ambition when he went to court was to remove the ‘inaccurate’ data; he had, apparently, no interest in *why* Google’s search algorithm continued to put long outdated results at the top of its rankings (even though arguably this was inexplicable given the high dimensionality of Google’s *PageRank* unsupervised learning system). A similar desire for action over explanation, can be seen in the various European ‘autocomplete defamation’ cases.⁴⁶⁰

⁴⁵⁷ *Google Spain* (n 405).

⁴⁵⁸ See generally Ausloos (n 303).

⁴⁵⁹ GDPR, art 17.

⁴⁶⁰ See generally Uta Kohl, ‘Google: the rise and rise of online intermediaries in the governance of the Internet and beyond (Part 2)’ (2013) 21(2) *International Journal of Law and Information Technology* 187 DOI: 10.1093/ijlit/eat004; Jones (n 129).

Explanations are not guaranteed to relieve or redress emotional or economic damage in all cases. They can more commonly allow *developers* not to make the same mistake again, when aimed at them. There are, of course, cases where explanations may tangibly enable decisions to be overturned: credit refusal issued by a machine, automated decisions to wrongfully refuse bail to a minority or welfare to someone with certain medical symptoms—these are obviously important social redresses—but explanations will not help or empower in all cases. Of course, if an individual explanation noted that race was a determining factor in a particular decision, then this would be easy grounds to complain. Yet this situation, where a decision can be seen in isolation to be unfair, is a parody ‘smoking gun’ example which we would hope never to happen. Firstly, it seems likely that if the organisation is so clearly rotten below the surface that any number of policy approaches might have unearthed problematic aspects. The cases which are more commonly problematic seem more nuanced, where systemic biases that are difficult to identify or pinpoint are the source of entrenched biases that affect society at large. Explanation facilities aimed at individuals are not set up to uncover systemic biases such as indirect correlation with race, and contrary to some who claim their utility for this,⁴⁶¹ they do not come with any technical or mathematical assurances they will uncover or expose these on much more than a luck basis. Transparency alone seems unlikely to provide either redress or public trust in the face of institutionalised power.⁴⁶²

A useful warning about transparency-based remedies or safeguards can be taken from the history of consent in information privacy. Privacy scholars are already very familiar with the notion that consent, often regarded by lay audiences as the primary safeguard for control of personal data,⁴⁶³ has in the online world become a mere husk of its former self, often described as meaningless, illusory, or otherwise unfit for purpose.⁴⁶⁴ Why is this? Online consent is most often obtained by displaying a link to a privacy policy at the time of entry to or registration with a site, app or network, and

⁴⁶¹ Wachter, Mittelstadt and Russell (n 205).

⁴⁶² See Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information* (n 160) 212.

⁴⁶³ See Elizabeth Denham, ‘Consent is not the ‘silver bullet’ for GDPR compliance’ [2016] Information Commissioner’s Office Blog (<https://iconewsblog.org.uk/2017/08/16/consent-is-not-the-silver-bullet-for-gdpr-compliance/>) accessed , (a blog from the UK DPA publicly attempting to downplay the sole importance of consent in relation to the renewal of data protection law).

⁴⁶⁴ Fred H Cate and Viktor Mayer-Schönberger, ‘Notice and consent in a world of Big Data’ (2013) 3(2) International Data Privacy Law 67 DOI: 10/cvcv; Christopher Kuner, Fred H Cate, Christopher Millard and Dan Jerker B Svantesson, ‘The challenge of ‘Big Data’ for data protection’ (2012) 2(2) International Data Privacy Law 47 DOI: 10/cvcx; Solon Barocas and Helen Nissenbaum, ‘Big Data’s End Run around Anonymity and Consent’ in *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (Cambridge University Press 2014) DOI: 10/cxvb; Mireille Hildebrandt, ‘Who is profiling who? Invisible visibility’ in *Reinventing Data Protection?* (Springer 2009) DOI: 10/ft4fnz; Ira S Rubinstein, ‘Big Data: The End of Privacy or a New Beginning?’ (2013) 3(2) International Data Privacy Law 74 DOI: 10/cvcz; Daniel J Solove, ‘Introduction: Privacy Self-Management and the Consent Dilemma’ (2012) 126 Harvard Law Review 1880.

2. The Law of Machine Learning?

asking the user to accede to these terms and conditions by ticking a box. As there is no chance to negotiate and little evidence that the majority of users either read, understand or truly consider these conditions, it is hard to see how this consent is either ‘freely given, specific, informed and unambiguous’ despite these being conditions for valid consent under the GDPR.⁴⁶⁵ The GDPR *does* attempt to improve the quality of consent with some new measures such as the requirement that the data controller must be able to prove consent was given,⁴⁶⁶ that terms relating to consent in user contracts must be distinguishable from other matters, and written in ‘clear and plain language’;⁴⁶⁷ and that in determining if consent was given ‘freely’, account should be taken of whether the provision of the service was conditional on the provision of data not necessary to provide that service.⁴⁶⁸ I submit however that these changes are not major, and that much will depend on the willingness of EU DPAs to take complex, expensive and possibly unenforceable actions against major data organisations such as GAFA emanating from non-EU origins with non-EU law norms.⁴⁶⁹

Consent as an online institution in fact arguably no longer provides any semblance of informational self-determination but merely legitimises the extraction of personal data from unwitting data subjects. Findings popularised in behavioural economics draw attention to the likelihood that many users have a faulty understanding of the privacy risks involved, due to asymmetric access to information and hard-wired human failure to properly assess future, intangible and contingent risks.⁴⁷⁰ Even in the real rather than online world, consent is manipulated by those such as employers or insurers whose business models and practices can exert pressures that render ‘free’ consent imaginary.⁴⁷¹ Even if we do posit in a rather utopian way that consent can be

⁴⁶⁵ GDPR, art 4(11).

⁴⁶⁶ GDPR, art 7(1).

⁴⁶⁷ GDPR, art 7(2).

⁴⁶⁸ GDPR, art 7(4).

⁴⁶⁹ The Common Statement of 5 DPAs is certainly an interesting first shot over the bows. See ‘Common Statement by the Contact Group of the Data Protection Authorities of The Netherlands, France, Spain, Hamburg and Belgium’ (CNIL, May 2017) (<https://www.cnil.fr/fr/node/23602>) announcing a number of privacy breaches by Facebook, one issue being that the company ‘uses sensitive personal data from users without their explicit consent. For example, data relating to sexual preferences were used to show targeted advertisements’. (noted specifically by the *Autoriteit Persoonsgegevens*, the DPA of the Netherlands). It is not said if that data was created algorithmically or existed as a user input. In addition, the author of this thesis has triggered several investigations against these firms, in particular against Google and its advertising ecosystem (Douglas Busvine, ‘Mozilla co-founder’s Brave files adtech complaint against Google’ (*Reuters*, 12th September 2018) (<https://perma.cc/FRC8-6L8W>)), Facebook (Rebecca Hill, ‘Chap asks Facebook for data on his web activity, Facebook says no, now watchdog’s on the case’ (*The Register*, 24th August 2018) (<https://perma.cc/AT5V-VSEU>)) and Twitter (Sara Merken, ‘Irish Twitter Probe seen as Test Case for EU Privacy Rules’ (*Bloomberg Law*, 18th October 2018) (<https://perma.cc/M83P-HX27>); Sean Pollock, ‘Twitter faces investigation by privacy watchdog over user tracking’ *The Sunday Times* (London, 21st October 2018)).

⁴⁷⁰ Tversky and Kahneman (n 166).

⁴⁷¹ See as a result the discussion by the A29WP of when consent is a valid grounds for processing in an employment context. Article 29 Data Protection Working Party, *Opinion 2/2017 on data processing at work*

given once to a data controller in a free and informed way (something recently pushed back against by AG Bobek in the opinion on the pending CJEU case *Fashion ID*⁴⁷²), constant vigilance will be needed as privacy policies and practices change frequently. It is unreasonable and increasingly unsustainable to abide by the neo-liberal paradigm and expect ordinary users to manage their own privacy via consent in the world of on-line dependence and what has been poetically called ‘bastard data’—where data ‘have become fertile and have bastard offspring that create new challenges that go far beyond what society previously (and, unfortunately, still) considered to be ‘privacy’.⁴⁷³

This state of affairs is can be summarised as the *transparency fallacy*—a term encapsulating the chasm between envisaged, ideal uses of transparency remedies, often portrayed as a panacea to various systemic ills, and the empirical roles they would or could (or indeed, fail to) play in individuals’ own lives. This empirical chasm has been under-researched by scholars of data and the law⁴⁷⁴—but without an understanding of how remedies are or are not useful to those affected, emerging law might not just fail to keep up with or anticipate technological change, but fail to anticipate its own utility. It is commonly noted by scholars that there are not just technical barriers to understanding machine learning systems, but skill-related and cognitive barriers too.⁴⁷⁵ Not considering these might give the semblance of having solved a problem, particularly as, similarly to FOI rights, journalists, civil society and politicians are some of the main users of transparency laws for purposes of investigation and holding other actors to account,⁴⁷⁶ and may therefore develop a misleading picture of the penetration of knowledge of such rights in society more broadly.

2.3.2. Weak collective provisions

Data protection is a framework with a heavy emphasis on individual rights. As a paradigm based on human rights, it has not typically been seen to contemplate remedies oriented around groups—or indeed around non-living persons such as corpora-

(WP 249) (2017).

⁴⁷² *Fashion ID* (n 407).

⁴⁷³ Joe McNamee, ‘Is Privacy Still Relevant in a World of Bastard Data?’ (*EDRI editorial*) (<https://edri.org/endoritorial-is-privacy-still-relevant-in-a-world-of-bastard-data>).

⁴⁷⁴ See generally Lee A Bygrave, ‘Legal Scholarship on Data Protection: Future Challenges and Directions’ in Cécile de Terwangne, Elise Degrave, Séverine Dusollier and Robert Queck (eds), *Liber amicorum Yves Poullet/Essays in honour of Yves Poullet* (Bruylant 2017) (<https://ssrn.com/abstract=3076747>).

⁴⁷⁵ Burrell (n 194). On challenges understanding machine learning systems, see section 1.6.1, p. 54..

⁴⁷⁶ See e.g. Daniel Swallow and Gabrielle Bourke, *The Freedom of Information Act and Higher Education: The experience of FOI officers in the UK* (The Constitution Unit, University College London June 2012) (<https://www.ucl.ac.uk/constitution-unit/research/foi/foi-universities/he-foi-officers-survey-report.pdf>) (a survey on the most common users of Freedom of Information rights by officers at higher education institutions in the UK).

2. The Law of Machine Learning?

tions or the deceased.⁴⁷⁷

Firstly, there are core limitations around who can trigger rights used to investigate potential infringements of the law. In line with the discussion of the *transparency fallacy*, it is heavily burdensome for individuals to exercise their rights under data protection and act as citizen-detectives. Data processing is near ubiquitous in modern society, and the time, effort, expertise and awareness needed to make informed choices is not likely to be possible for the majority of citizens, let alone vulnerable individuals who might be most at risk from granular profiling techniques. This is even truer perhaps in the EU, where consumers are perhaps on the whole far less prepared and empowered to litigate than elsewhere, particularly than the US.⁴⁷⁸ Individuals are further hampered in meaningfully attaining civil justice by a general prejudice against contingency lawyering combined with dwindling levels of civil legal aid.⁴⁷⁹ In cases where individual burden seems disproportionate, collective or class action provisions—such as those in consumer law—seem effective ways to alleviate these issues. The UK and many other EU nations have no generic system of class actions, although a variety of class-action-flavoured solutions exist internationally, such as the Netherlands' rules on *collective settlement*.⁴⁸⁰ Although this has been viewed as a problem for many years, attempts to solve it on an EU wide basis have repeatedly stalled.

The GDPR *does* contain some limited collective action provisions that go beyond what was specified in the DPD. These take two main forms.

The first part of the collective action provisions in the GDPR is mandatory for all member states.⁴⁸¹ In effect, these provisions allow individuals to explicitly delegate certain rights to a non-profit entity. These are *not* the rights discussed above such as information rights or objection rights however, but instead are limited to the *right to lodge a complaint with a supervisory authority*, the *right to an effective judicial remedy against a supervisory authority*, the *right to an effective judicial remedy against a controller or processor* and the *right to compensation* for material and non-material damage. Effectively, these allow individuals the right to complain, the right to take a DPA, controller or processor to court, and a right to damages. They gain a potential collective dimension when many individuals delegate their rights at the same time to the same entity. Yet

⁴⁷⁷ See Lilian Edwards and Edina Harbinja, 'Protecting Post-Mortem Privacy: Reconsidering the Privacy Interests of the Deceased in a Digital World' (2013) 32 *Cardozo Arts & Ent L J* 83 DOI: 10/gfhxns, 113. See also GDPR, recital 27.

⁴⁷⁸ Although such comparisons should be made with care due to differences across sectors and specific claims; cf Eoin Quill and Raymond J Friel (eds), *Damages and Compensation Culture: Comparative Perspectives* (Bloomsbury Publishing 2016).

⁴⁷⁹ The Law Society of England and Wales, *Legal aid deserts in England and Wales* (2016) (<https://perma.cc/268A-5TJM>).

⁴⁸⁰ Stacie I Strong, 'From Class to Collective: The De-Americanization of Class Arbitration' (2010) 26(4) *Arbitration International* 493 DOI: 10/gfr9hf, 505–506.

⁴⁸¹ GDPR, art 80(1).

while this compels member states to allow the delegation of these rights, it does not by default permit the *combination* of legal proceedings, at least in the UK. In the UK's Data Protection Act 2018, the right to allow such combination was not explicitly added, but the Secretary of State would be allowed to make such combination possible through future regulations.⁴⁸²

The second part of the collective action provisions is non-mandatory⁴⁸³—that is to say, it does not have direct effect, and it is up to Member States to pass their own provisions, which the regulation provides for and allows. It is most familiar to UK readers as a variant on the *super-complaint* mechanism. The UK is somewhat unusual among European countries in having legislation empowering civil society organisation to make fast-tracked complaints in some sectors, such as regarding adverse effects of the operation of a market for consumer interests.⁴⁸⁴ Three main areas with active super-complaint mechanisms currently exist in the UK: consumer protection, financial services, and payment services.⁴⁸⁵ The typical set-up of such a provision requires a regulator to respond to a civil society complaint (from a list of specified organisations⁴⁸⁶) within a certain time-frame, usually 90 days, stating whether they will take action on a super-complaint and their reasons for doing so or refusing to. Despite perhaps being the country with the most experience of such mechanisms, the UK Government declined to add a derogation concerning this provision in the implementing legislation for the GDPR, conceding only (after pressure in the House of Lords) to a review two years after Royal Assent.⁴⁸⁷ While it seems unlikely that the Government was concerned about the ability of civil society to alert regulators of potential infringements of the law (as the traditional super-complaint mechanisms do), there is

⁴⁸² DPA 2018, s 188.

⁴⁸³ GDPR, art 80(2).

⁴⁸⁴ See generally Colin Scott, 'Enforcing Consumer Protection Laws' in Geraint Howells, Iain Ramsay and Thomas Wilhelmsson (eds), *Handbook of Research on International Consumer Law, Second Edition* (Edward Elgar 2018) DOI: 10.4337/9781785368219.

⁴⁸⁵ Enterprise Act 2002, s 11, where a consumer body designated by the Secretary of State may make a complaint to the Competitions and Markets Authority (or a selection of other authorities, see The Enterprise Act 2002 (Super-complaints to Regulators) Order 2003, SI 2003/1368) 'that any feature, or combination of features, of a market in the United Kingdom for goods or services is or appears to be significantly harming the interests of consumers'; Financial Services and Markets Act 2000, s 234C, where a 'designated consumer body may make a complaint to the [Financial Conduct Authority] that a feature, or combination of features, of a market in the United Kingdom for financial services is, or appears to be, significantly damaging the interests of consumers'; Financial Services (Banking Reform) Act 2013, s 68, where a 'designated representative body may make a complaint to the Payment Systems Regulator that a feature, or combination of features, of a market in the United Kingdom for services provided by payment systems is, or appears to be, significantly damaging the interests of those who use, or are likely to use, those services'.

⁴⁸⁶ For example, *Consumer Focus* (better known as *Which?*) and Citizens Advice are currently permitted to complain under UK consumer law.

⁴⁸⁷ For the debate, see HL Deb 10th January 2018, vol 788, 286; for the provision added post-debate, see DPA 2018, s 189.

2. The Law of Machine Learning?

a stark difference between the GDPR provision and super-complaints insofar as this would give civil society organisations rights to bring court cases against both DPAs and against data controllers and/or processors. This would effectively add a form of class action which the UK is largely unfamiliar with into the legal system: hence the cautious description of these GDPR ‘super-complaints’ by the Government at the time of the debate of the Data Protection Act 2018 as entailing ‘risks and potential pitfalls’.⁴⁸⁸ It is worth noting however, that these super-complaint-like mechanisms in Article 80(2) *do not* include the *right to compensation* for material and non-material damage, which can still only be triggered by an individual either taking a case themselves or explicitly delegating that right to a non-profit entity within the meaning of Article 80(1).

Interestingly, a new proposal for a Directive on representative actions for the protection of the collective interests of consumers has recently been put forward by the European Commission.⁴⁸⁹ This would build on existing legal structures and forms for only eight current EU Member States, while introducing a completely new and strange legal mechanism to the legal traditions and practices of the twenty others.⁴⁹⁰ The proposal suggests that it will cover data protection issues within its remit—how this mechanism interplays with and extends the GDPR mechanisms, assuming it will pass, is yet to be seen.⁴⁹¹ With regards to the UK, it seems unlikely this instrument would enter UK law were the UK to leave the European Union.

Do these provisions and processes mean that the individualistic problem of data protection is solved? Simply put, no. The core reason for this is that such proposals and initiatives might welcome support *once infringements have been identified*, but provide little-to-no support in the identification of such challenges, nor in the way that individuals are supposed to manage their own risks concerning data processing within the data protection framework.

As discussed, the *identification* of infringements of the law is one of the serious challenges for data protection law, and one where the policy approach has traditionally been to rely on a system of policing supported by individuals using their own rights.⁴⁹² Neither the approach in Article 80 nor any proposed extension of representative actions by the European Commission tackles the tricky fact that this theory of enforce-

⁴⁸⁸ HL Deb 10th January 2018, vol 788, 286.

⁴⁸⁹ Commission, ‘Proposal for a Directive of the European Parliament and of the Council on representative actions for the protection of the collective interests of consumers, and repealing Directive 2009/22/EC’ COM (2018) 184 final.

⁴⁹⁰ Ignasi Guardans, ‘A New (Sort of) Class Action in Protection of European Consumers’ (*Lexology*, 25th April 2018) (<https://perma.cc/X2L2-9J4Y>).

⁴⁹¹ Note that the European Data Protection Supervisor (EDPS) has already highlighted some tensions between this proposal and data protection law as it currently stands. See European Data Protection Supervisor, *Opinion 8/2018 on the legislative package “A New Deal for Consumers”* (EDPS 2018).

⁴⁹² See above section 2.2.2.1, p. 107.

ment is fatally undermined at present by the same forces at play as in the transparency fallacy. The rights they cover surround highlighting known violations and seeking their remedies from regulators or courts, *not* from i) using individuals' rights to better control their data in line with their wishes or ii) using such rights to uncover unknown violations in the first place. A non-profit within the meaning of the collective action provisions of the GDPR cannot object on your behalf, cannot access data or trigger transparency provisions on your behalf, and cannot manage consents on your behalf. Were they to be able to, for example, object on behalf of thousands of hundreds of thousands of people to the creation of a machine learning model they believed to be harmful, they might turn 'voting with your feet' into an actionable reality in the world of data governance.⁴⁹³

Some rights could in theory be delegated with ease. Objection—where any processing based on *legitimate interests* or *public task* (as opposed to eg consent or contract) can be contested and subject to a balancing test—is a somewhat simple right to delegate, with historical precedent in 'Do Not Call' registries in the US⁴⁹⁴ or marketing call blacklist schemes such as *BT Call Protect* in the UK. Furthermore, there are some hooks in the legislation which might allow the *de facto* delegation of this right through technical means. In certain cases, the data subject should be able to 'exercise his or her right to object by automated means using technical specifications'—a privilege not afforded to any other right in the GDPR.⁴⁹⁵ This enters the fraught world of the politics of binding Do Not Track (DNT) signals, which are a matter of current debate in the negotiations around the draft ePrivacy Regulation.⁴⁹⁶ These might hold promise if some standardisation and definitional challenges can be overcome.⁴⁹⁷

Access however remains a prerequisite for civil society organisations to understand what should be objected to, and is considerably trickier to delegate due to the large privacy challenges involved, and the undesirability of centralising sensitive personal data (as opposed to, for example, a list of people and their objection preferences) inside civil society organisations that might not be empowered to secure it. Where does

⁴⁹³ See generally René LP Mahieu, Hadi Asghari and Michel van Eeten, 'Collectively Exercising the Right of Access: Individual Effort, Societal Effect' (2018) 7(3) *Internet Policy Rev.* DOI: 10/cwd8.

⁴⁹⁴ Irene Kamara and Eleni Kosta, 'Do Not Track Initiatives: Regaining the Lost User Control' (2016) 6(4) *International Data Privacy Law* 276 DOI: 10/gdxwds, 280.

⁴⁹⁵ GDPR, art 21(5).

⁴⁹⁶ Commission, 'Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications)' COM(2018) 640 final; Committee on Civil Liberties, Justice and Home Affairs, *Report on the proposal for a regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications)* (European Parliament 2017).

⁴⁹⁷ See Kamara and Kosta (n 494) 287–289.

2. The Law of Machine Learning?

collective access fail currently?

The regulators themselves have powers ‘to obtain, from the controller and the processor, access to all personal data and to all information necessary for the performance of [the DPA’s] tasks’.⁴⁹⁸ Parts of the UK parliamentary apparatus, such as Select Committees, have also recently made use of somewhat arcane powers in order to try to uncover breaches of data protection. The UK Commons Select Committee for Digital, Culture, Media and Sport (overseeing the Department with responsibility for data protection legislation) has arguably pushed these powers to the limit, using (or at least threatening) a Speaker’s Warrant to effectively force the disclosure (via a plaintiff) of documents from Facebook concerning its data sharing operations which had been obtained through civil discovery and sealed by the San Mateo Superior Court, a state court of California.⁴⁹⁹

When issues get taken to court—which itself presumes there is enough information to suspect a breach that would lead a civil society organisation to take the risk of litigating—there are further challenges in relation to investigation and transparency around issues affecting groups. When courts do get involved, eg where judicial review is invoked, they seem likely to continue a current trend of being reluctant to order disclosure of the code of models even within a secure setting, primarily due to concerns around trade secrets or sensitive information. While this has been seen more explicitly in the US,⁵⁰⁰ in the UK, there appears to be no reported case where a court has ordered disclosure of the source code of a decision support system to litigants. Interestingly in at least one case,⁵⁰¹ the output of a conventional though complex automated decision support system was doubted in respect to its value without more information as to how it was generated. The system in question calculated one factor (economic rent) to feed into a compensation valuation in cases of compulsory acquisition of land. The court was disturbed at the lack of evidence it received as to exactly how this calculation had been done but, interestingly, did not seem interested to find out more but rather to exclude it from influence. It seems quite likely that courts will be reluctant to become activists about disclosures of source code, let alone algorithmic training sets and models, until they feel more confident of their ability to comprehend and use such evidence—which may take some time.

Individuals can of course request their own data as well as metadata about the processing activities that relate to them, but this would still need to be combined with

⁴⁹⁸ GDPR, art 58(1)(e)

⁴⁹⁹ Carole Cadwalladr, ‘Parliament seizes cache of Facebook internal papers’ (*The Observer*, 24th November 2018) (<https://perma.cc/T7TH-U8AF>).

⁵⁰⁰ See eg, in the US, *Viacom Intern. Inc. v. YouTube, Inc.*, 718 F. Supp. 2d 514 (S.D.N.Y. 2010) (where the Court refused to order the hand over of Google’s proprietary search algorithm).

⁵⁰¹ *Northern Metco Estates Ltd v Perth and Kinross DC* 1993 SLT (Lands Tr) 28.

other users to give a picture of how processing is affecting groups and society rather than just them alone. With only a single instance of data, very little can be said about the impact of personalisation or the possibility of discrimination or manipulation, as there is no reference point for comparison. To make such a reference point would be possible were users to combine the data they got from different regulators, yet automated subject access rights are highly challenging in practice with several legal, social and technical barriers.⁵⁰²

After the powers of regulators and Parliament, there is a precipitous gap of oversight, which has to be carried out by journalists, researchers and civil society organisations without any special access to data or systems in the private sector. FOI rules provide transparency in the public sector, and can dodge privacy concerns as the information received is supposed to be unaffected by who asks for it.⁵⁰³ Where an information request is refused, stronger powers of regulators kick in. The Information Commissioner can oversee these rights by issuing a powerful information notice to FOI-able bodies,⁵⁰⁴ usually followed by a determination of whether the grounds for refusal was valid. Yet these rules do not apply to the private sector—in the UK, controversially, they do not even apply to public contractors.⁵⁰⁵ Researchers and journalists have resorted to more devious means to understand automated systems, such as by creating automated accounts and using bots to detect discrimination online.⁵⁰⁶ Yet in some contexts, the use of tools to investigate platforms to better understand their data practices, such as bots, might themselves be in breach of the law, particularly if their use is construed as hacking. In the US, the potential for studies looking at algorithmic discrimination to breach the Computer Fraud and Abuse Act (CFAA),⁵⁰⁷ has led to a lawsuit testing and challenging the constitutionality of these provisions.⁵⁰⁸ A related argument can

⁵⁰² See generally Michael Veale, Lilian Edwards, David Evers, Tristan Henderson, Christopher Millard and Barbara Staudt Lerner, 'Automating Data Rights' in David Evers, Christopher Millard, Margo Seltzer and Jatinder Singh (eds), *Towards Accountable Systems (Dagstuhl Seminar 18181)* (Dagstuhl Reports 8(4), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik 2018) DOI: 10/gffngz.

⁵⁰³ Although some journalists who would plan to publicise the outcomes suspect they receive less than less-connected members of the public might have. The author has heard this personally, and it is also reported in the research literature, eg Michael Hunt, 'Local Government, Freedom of Expression and Participation' in Richard A Chapman and Michael Hunt (eds), *Freedom of information: Local government and accountability* (Ashgate Publishing, Ltd 2010) 51.

⁵⁰⁴ Freedom of Information Act 2000, s 51.

⁵⁰⁵ See eg recommendations for change from the ICO Information Commissioner's Office, *Transparency in outsourcing: a roadmap* (ICO 2015) (<https://ico.org.uk/media/1043531/transparency-in-outsourcing-roadmap.pdf>) ¶¶ 54–55 and a corresponding Private Members' Bill currently in the parliamentary pipeline: Freedom of Information (Extension) Bill HC Bill (2017–19) [23].

⁵⁰⁶ See eg Datta, Tschantz and Datta (n 104); see generally Christian Sandvig, Kevin Hamilton, Karrie Karahalios and Cedric Langbort, 'Auditing algorithms: Research methods for detecting discrimination on internet platforms' in *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* (Seattle, WA, 2014) (<https://perma.cc/8JKK-FMUV>).

⁵⁰⁷ 18 U.S.C. §1030 *et seq.*

⁵⁰⁸ See *Sandvig v. Sessions*, No. 16-1368 (JDB), 2018 WL 1568881 (D.D.C. Mar. 30, 2018).

2. The Law of Machine Learning?

be made in relation to the ‘chilling effect’ of UK computer security law on vulnerability discovery.⁵⁰⁹

Given that personal data transparency relates to the asker and that anyone *but* the asker allowed to access personal data might risk a privacy breach, is this weakness of data protection inevitable? No. The kind of transparency (thought to be) provided through the automated decision provisions in data protection *could* be meaningfully aggregated and scaled up in a way which preserved individual privacy. All that would be at stake would be the individual decisions and policies a firm made. While arguably this could be a trade secret challenge, we have seen recently an obligation for releasing comparably sensitive materials in relation to the gender pay gap publishing obligations, which require employers of over 250 workers publish the male/female difference between average hourly rate of pay, average bonus and proportions of those receiving them, and employees in each quartile pay band of the workforce.⁵¹⁰ Similarly, reporting on aggregate factors such as the demographics machine learning systems were trained on and applied to, performance metrics over time and on subgroups—the ‘model-centric explanations’ discussed earlier⁵¹¹— seem wholly amenable to open up to either general publishing or the ability for certain entities to request on the behalf of affected societal groups.

There are also challenges in the GDPR relating to the difference between group and individual impacts. Return for a moment to the *Sweeney Search* algorithmic war story, where a Harvard professor found, when she Google searched her name, that it was linked, seemingly by proxy of her membership of an ethnic group, to impugned criminality.⁵¹² Further assume that a search system constituted a ‘decision’ based solely on automated processing.⁵¹³ Whether or not the protections of the GDPR in relation to such decision trigger then largely depends on whether this decision had ‘legal [...] or similarly [significant]’ effect.⁵¹⁴ Clearly there was no effect on Sweeney’s legal status (which implies changes to public law status such as being classified as a US citizen, or private law effects such as having capacity to make a will). But did such a decision have a *significant* effect on Sweeney? The most obvious takeaway is that a *racial group* was affected by an assumption of above average criminality, and she was part of that group, which although a familiar formulation in discrimination laws, takes us to somewhere very different from the individual subject-focused rights usually gran-

⁵⁰⁹ Audrey Guinchard, ‘The Computer Misuse Act 1990 to Support Vulnerability Research? Proposal for a Defence for Hacking as a Strategy in the Fight against Cybercrime.’ (2018) 2(2) *Journal of Information Rights, Policy and Practice* DOI: 10/gfscft.

⁵¹⁰ The Equality Act 2010 (Gender Pay Gap Information) Regulations 2017, SI 2017/172.

⁵¹¹ See section 1.6.1, p. 54.

⁵¹² See section 1.4.2, p. 42.

⁵¹³ See section 2.2.1.2, p. 101.

⁵¹⁴ GDPR, art 22(1).

ted by data protection and the GDPR. Even if we accept an impact on Sweeney as an individual constructed through group membership, its significance on her as an individual could be doubted. She did after all merely have sight of an advert which she was not compelled to click on, and which could even have been hidden using an ad blocker.

In the *Jew Watch* war story (where an anti-Semitic website was the top result when an individual searched the term ‘Jew’),⁵¹⁵ it is even harder to say a ‘decision’ was made affecting any one individual—even though its significance seems high, given the prominence of the result. Given the complexity of the search algorithms involved, dependent not only on variables derived from the searcher but also the general search environment, it is very hard to predict a particular ranking of sites being shown to a particular user in advance. Furthermore, the searcher might not themselves be of the class affected,⁵¹⁶ and so their ability to ‘raise the flag’ would be hampered as a result.

Some mechanisms are in place in Article 22 to cope with protected groups. Effectively, for private actors, where decisions are based on special category data such as that revealing race or religion, as arguably the case in *Sweeney Search*, the GDPR probably required that she had given that data to Google by explicit consent.⁵¹⁷ If that was so, she could potentially claim under Article 22(4) the ‘right to an explanation’ of how the advertising delivery algorithm had worked, or ‘meaningful information’ under Article 15.⁵¹⁸ Yet there is a *Catch-22* situation at play here. Was the decision based on race? Was it not more likely instead based on a multiplicity of ‘ordinary’ information that Sweeney provided as signals to the ranking algorithm, plus signals from the rest of the group which together might statistically proxy race? Perhaps it was based on information the advertiser provided to Google—trigger names or keywords, for example? To operationalise Article 22, including its related transparency provisions, a data subject needs to know the relevant input variables of a system—which itself may require access to something resembling the algorithmic explanation that is being sought through this whole process. This takes us back to the investigative problems above. Few companies will want to admit they are making automated decisions that require explicit, separate consent, and instead will want to justify them in different ways (such as contract, or by claiming they are not solely automated). A system which requires data controllers to

⁵¹⁵ See section 1.4.2, p. 42.

⁵¹⁶ This might be exacerbated in this case by the noted tendency for members of this class to avoid calling themselves ‘Jews’ and instead refer to themselves as ‘Jewish people’ or similar. See Mark Oppenheimer, ‘Reclaiming ‘Jew’’ (*The New York Times*, 22nd April 2017) (<https://perma.cc/8QM2-DHDB>).

⁵¹⁷ GDPR, art 9(2); GDPR, art 22(4).

⁵¹⁸ While scholars have argued the right to an explanation, in recitals, would not be binding (see eg Wachter, Mittelstadt and Floridi (n 201)), interestingly there is no list of safeguards for the case where special category data is used in such decisions (GDPR, art 22(4)), meaning the recitals are the only place to look, and that decisions based on such data might have more of a claim to a binding explanation right. See further Edwards and Veale, ‘Slave to the Algorithm?’ (n 79).

2. The Law of Machine Learning?

make this call themselves in ways that are not practically accountable to data subjects seems doomed to provide little protection at all.

Even *if* i) the conditions for Article 22 were engaged and ii) investigative abilities shone light on the fact Article 22 *should* be engaged and held a data controller to do so in practice, the *remedies provided* do little to practically aid groups and structural challenges caused by machine learning systems. The basic remedy, of a ‘human-in-the-loop’, does little to ensure fast-moving systems do not harm vulnerable groups to begin with, that they do not sow division or disadvantage, that they are effectively monitored and held to account. Indeed, at scale in many areas (such as upload filters) a human-in-the-loop is arguably a pipe dream rather than a feasible remedy—at worst incentivising the creation of an underpaid digital underclass, with little stability or progression, based in the cheapest and most precarious parts of the world.

2.4. Beyond individualism

While data protection does suffer from its individualistic focus, particularly in the area of automated decision-making, this need not cripple it irreparably. I now want to highlight several useful parts of the GDPR—foundations or building blocks, rather than fully-formed provisions—which might serve as useful seeds for more robust, sustainable and useful governance of machine learning within the data protection regime.

2.4.1. Purpose limitation

Firstly—and something which is often underemphasised—is the requirement to have a lawful basis and a specified purpose for any training of machine learning systems utilising personal data. Like all processing of personal data, the building of machine learning models needs a ground drawn from Article 6(1). This ground—such as consent, legitimate interest, or public task—needs to be tied to a *purpose*. As it stands, firms in practice often draw their purposes widely—to improve product and service offering, for example. Going by the previous guidance of the A29WP, such vague purposes are not lawful. In a 2013 opinion, the A29WP noted that ‘a purpose that is vague or general, such as for instance ‘improving users’ experience’, ‘marketing purposes’, ‘IT-security purposes’ or ‘future research’ will—without more detail—usually not meet the criteria of being ‘specific’⁵¹⁹ Such an approach is a powerful regulatory tool in light of increasing trends to collect data with no clear, or at least announced, purpose—and an underexplored one, particularly in light of the way that to date, DPAs have ‘shied away

⁵¹⁹ Article 29 Data Protection Working Party, *Opinion 03/2013 on purpose limitation (wp203)* (2013) 16.

from taking decisive steps towards [purpose limitation's] enforcement'.⁵²⁰ It is likely also for this reason that industry has been fervently opposed to purpose limitation taking a more central role in data protection. In a Microsoft-sponsored report, academics from the Oxford Internet Institute and Indiana University claimed that while purpose limitation 'may have been feasible in 1980, it does little to protect individuals today',⁵²¹ omitting it from their proposed revisions to data protection law (during the debates around the GDPR).

In many ways, the principle of purpose limitation is akin to the principle of legal certainty, that justice is not possible unless those involved are capable of foreseeing the implications of their actions. 'On one side of the coin,' writes Mireille Hildebrandt, 'the notion of purpose constitutes and configures the role of the data controller, by forcing it to determine a purpose for each processing operation. On the other side of the same coin, the need to determine the purpose creates awareness of the liability for usage beyond the specified purpose. In this manner the purpose ensures that sharing data with a controller has reasonably foreseeable consequences.'⁵²²

The challenge running through this thesis is that machine learning systems might have consequences which were not, *prima facie*, foreseeable. These include issues around fairness and privacy at individual and societal levels already discussed, but might even extend to broader failures. If we see algorithmic systems as complex and adaptive, fast-moving and with limited lag time between cause and effect, then the 'normal accidents' hypothesis from the sociology of engineering systems suggests that we should *expect* adverse consequences which have causal reasoning difficult to pre-empt.⁵²³

As a consequence, I argue that it is possible to usefully extend purpose limitation in the direction of safeguards that attempt to mitigate or limit these risks. This is drawn from the general approach advocated by the A29WP around specifying purposes such that data subjects can grapple with them: the *layered notice*.⁵²⁴ The layered notice allows individuals to 'drill-down' into as much (or as little) detail as they require. If the purposes are high-risk—they come with the risks of the algorithmic war-stories already elaborated upon—then purposes for which data are used, whether it using or training models, should be elaborated upon, potentially extensively. In the same vein, lawful bases and purposes are *granular* in data protection. Take-it-or-leave-it of-

⁵²⁰ Judith Rauhofer, 'Of Men and Mice: Should the EU Data Protection Authorities' Reaction to Google's New Privacy Policy Raise Concern for the Future of the Purpose Limitation Principle' (2015) 1 European Data Protection Law Review (EDPL) 5 DOI: 10/gfsh4s, 15.

⁵²¹ Fred H Cate, Peter Cullen and Victor Mayer-Schönberger, *Data Protection Principles for the 21st Century: Revising the 1980 OECD Guidelines* (, Oxford Internet Institute 2014) (<https://perma.cc/G5HL-VUPL>) 6.

⁵²² Hildebrandt, *Smart technologies and the End(s) of Law* (n 68) 205.

⁵²³ For an analogy in industrial systems, see Perrow (n 375).

⁵²⁴ Article 29 Data Protection Working Party, *Opinion 03/2013 on purpose limitation (wp203)* (n 519).

2. The Law of Machine Learning?

fers are broadly disallowed, and individuals can refuse consent or object to particular purposes.⁵²⁵ This is likely to be examined in future cases before the Court, as well as connected to debates around the proposed ePrivacy Regulation in the particular contexts of web tracking and connected devices,⁵²⁶ and we may have to wait for complete clarity.

Purpose limitation could apply to two prototypical data subjects. The first has their data collected to *build* a model, while the second has their data collected to *query* one (and, perhaps, apply its results). In both cases, there is room to expand the interpretation of data protection law to state not just that models *are* being trained or queried, but *which* models, to what ends, and with what safeguards. For those involved in training the models, this would give them the ability to better control what insights from their data would be used for (the subject of section 3.2 in the next chapter, with further elaboration on reasoning described therein). For those involved with querying them, this bolsters their transparency rights described above.

Purpose limitation does not fall fully within the same individualistic trap as many other aspects of data protection law previously examined for a number of reasons. Firstly, purposes have to be provided for all processing, before processing (including collection) begins. In practice, this is often done publicly to lower communication overheads: this is the privacy policy found at the bottom of websites the world over, which in the US has a contractual flavour, but in Europe primarily exists to fulfil the requirements of Articles 13–14 of the GDPR on transparency requirements. Such privacy policies can be accessed by data subjects and civil society organisations alike, and, were model information to be provided therein, can form a useful starting point for further investigation and potentially enforcement action. Such analysis might even benefit from automation and web-scraping technologies which themselves might be enhanced by machine learning.

Purpose limitation is an important tool in machine learning governance beyond individualism after such information has been scraped, as then it may be possible, with the help of digital tools, for individuals to effectively broadcast detailed information on what they consent to and what they do not (or what they object to and what they do not) via automated means. If purposes are specific enough, this opens the door

⁵²⁵ The GDPR does not completely outlaw take-it-or-leave-it offers, but does specify that ‘utmost account’ be taken of any contract or service made conditional on consent of any processing not necessary for its provision. See GDPR, art 7(4) and more broadly, Frederik Zuiderveen Borgesius, Sanne Kruikemeier, Sophie C Boerman and Natali Helberger, ‘Tracking Walls, Take-It-Or-Leave-It Choices, the GDPR, and the ePrivacy Regulation’ (2017) 3(3) European Data Protection Law Review 353 DOI: 10/gfsh4x.

⁵²⁶ The EDPS for example recommends further strengthening language that would prevent take-it-or-leave-it choices in the proposed ePrivacy regulation, see European Data Protection Supervisor, *Opinion 6/2017: EDPS Opinion on the Proposal for a Regulation on Privacy and Electronic Communications (ePrivacy Regulation)* (EDPS 2017) 16–18.

for them to be individually scrutinised and held to account—or subject to further information rights, such as those in Article 15—by data subject or on their behalf by civil society organisations. Individuals might subscribe to a consent list that automatically places an objection to building emotion recognition systems that are sold commercially, for example, for concerns over how they would be used. Particular policies (such as their commercial use) could force privacy policies to display such information, if the civil society list in question assumed an objection unless all the information they demanded was included. Thus, such granular control could, in theory, be used to set standards in a relatively coercive manner. If machine learning systems are not trained with regard to fairness; if they are not made openly available; or if they are used in ways which are not monitored and evaluated in public fora, then this could be grounds for individuals, through civil society organisations, to take punitive action by refusing use of their data. Such practices are already emerging today. The Electronic Frontier Foundation (EFF) are attempting to mandate a *Do Not Track* standard by blocking, insofar as is possible, tracking and other revenue-making functionality such as adverts specifically for firms which do not honour web browser signals or commands, using their plug-in *Privacy Badger*.⁵²⁷

2.4.2. Data Protection Impact Assessments

The GDPR specifically introduces a number of new provisions which do not confer individual rights but attempt to create an environment in which less ‘toxic’ or otherwise harmful automated systems will, or can be, built and deployed. Broadly, these ideas emerged from the research and practice around privacy by design (PbD)—using engineering methods to build privacy-aware or -friendly systems, starting from the beginning of the process rather than tacking it on at the end as an afterthought. They also emerge from a recognition that top-down regulation around data, as common, fluid and context-specific as it is, can only go so far—and controllers themselves must be involved in the design of better systems.

Data protection impact assessments (DPIAs) in particular have large implications for the design and deployment of algorithmic systems. Formerly known as privacy impact assessments (PIAs) (and still known as that in many parts of the world), in practice they had up until now primarily been taken up by public bodies, such as health trusts, and had remained a voluntary endeavour. In the GDPR, they became compulsory (in certain situations)⁵²⁸ under their new guise of the DPIA. The GDPR requires a

⁵²⁷ Electronic Frontier Foundation, ‘Do Not Track’ (*eff.org*, 31st December 2018) (<https://perma.cc/CF2M-YFZR>) (‘Privacy Badger offers good actors the option of having their content (including ads) unblocked if they adopt EFF’s DNT policy’).

⁵²⁸ GDPR, art 35.

2. The Law of Machine Learning?

DPIA where, in particular there is a ‘systematic and extensive evaluation of personal aspects relating to natural persons [...] based on automated processing, including profiling [...] and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person.’

As an illustration of how quickly this field is moving, at the time of writing a preliminary version of this argument,⁵²⁹ a co-author and I noted that ‘DPIAs are quite likely to become the required norm for algorithmic systems’⁵³⁰ and that ‘a DPIA will be an obligatory precursor for many [machine learning] systems with sizable anticipated risks or consequences for individuals or groups’.⁵³¹ Since then, the ICO has, in line with the GDPR,⁵³² created a list of ‘examples of processing ‘likely to result in high risk’ within the meaning of the DPIA provisions. Number one on that list is indeed and specifically, machine learning.⁵³³

Data protection impact assessments may fulfil some of the desires of model-centric explanations described previously: requiring controllers to outline any bias detection measures, monitoring and evaluation strategies, and mitigation approaches used to ensure that machine learning systems support the rights and freedoms of data subjects. Interestingly, the ‘rights and freedoms’ that the DPIA provisions concern appear not to be limited to those in data protection law, but those applicable more broadly to others, such as freedom of expression or anti-discrimination.⁵³⁴

DPIAs might also make notions of better *due process* around algorithmic systems more of a reality within private data controllers. Where consequential (administrative) decisions which might affect individuals, such as deprive them of their liberties, are taken, constraints on the rules that can be made in combination with procedural due process around the act of making the decision (and the systems involved) provides basic structural protection for individuals concerned. In relation to automated and/or data-driven systems, it has been argued that these due process norms are threatened.⁵³⁵ Some of these issues are those already discussed above in data pro-

⁵²⁹ Edwards and Veale, ‘Slave to the Algorithm?’ (n 79).

⁵³⁰ *ibid.*

⁵³¹ Lilian Edwards and Michael Veale, ‘Enslaving the Algorithm: From a “Right to an Explanation” to a “Right to Better Decisions”?’ (2018) 16(3) IEEE Security & Privacy 46 DOI: 10/gdz29v, 51.

⁵³² GDPR, art 35(5).

⁵³³ Information Commissioner’s Office, *Examples of processing ‘likely to result in high risk’* (ICO 2018) (<https://perma.cc/SRS6-M7WX>) accessed 28th December 2018.

⁵³⁴ See Article 29 Data Protection Working Party, *Statement on the role of a risk-based approach in data protection legal frameworks* (2014) 4; 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 (WP 248 rev.01)* (2017) 6 (‘the reference to “the rights and freedoms” of data subjects primarily concerns the rights to data protection and privacy but may also involve other fundamental rights such as freedom of speech, freedom of thought, freedom of movement, prohibition of discrimination, right to liberty, conscience and religion’).

⁵³⁵ See eg Danielle Keats Citron, ‘Technological due process’ (2008) 85 Washington University Law Review

tection law: that individuals are alerted of significant solely automated decisions with enough to time challenge or respond to them. Such mechanisms can also be seen in some administrative law systems, such as the French system, where the 2016 *Loi pour une république numérique* (Digital Republic Act)⁵³⁶ provides additional informational safeguards regarding ‘algorithmic treatment’ by the public sector.⁵³⁷ In these cases, the logic is that individuals should be aware of such systems with sufficient lead-time as to effectively respond to them, mitigating decisions that are based on incorrect information, misapplied rules, or other false premises.⁵³⁸

Yet one further important part of procedural integrity around rule-making are notice-and-comment periods when a rule has been changed or been implemented.⁵³⁹ As already seen, machine learning systems encode rules in subjective ways that often deserve oversight. In the public sector, notice-and-comment rules (also known as a duty to consult) are common. In the UK, while there is no omnibus duty to consult, one can be generated by statute or a common law duty upon a public authority to act fairly and in line with legitimate expectations.⁵⁴⁰ Regardless of the duty, which will differ by jurisdiction, administrative law fails to address consequential private decisions which might also benefit from a flavour of due process. DPIAs might address these to some degree, or provide a foundational for doing so. Indeed, they do come with a consultation requirement for stakeholders, albeit one which is highly qualified to restrict consultation to ‘where appropriate’ and ‘without prejudice to the protection of commercial or public interests or the security of the processing operations’.⁵⁴¹ This requirement, much stronger in the European Parliament text during the GDPR negotiation process but weakened in trialogue,⁵⁴² might still provide a foundation or hook for engaging stakeholders upstream in the development of automated systems in a way envisaged by proponents of better ‘technological due process’.⁵⁴³

One of the core problems with DPIAs however, and one that compounds the weak

1249; Keats Citron and Pasquale (n 417); Kate Crawford and Jason Schultz, ‘Big Data and due process: Toward a framework to redress predictive privacy harms’ (2014) 55 Boston College Law Review 93.

⁵³⁶ Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, art 4. (stating “[...] une décision individuelle prise sur le fondement d’un traitement algorithmique comporte une mention explicite en informant l’intéressé. Les règles définissant ce traitement ainsi que les principales caractéristiques de sa mise en œuvre sont communiquées par l’administration à l’intéressé s’il en fait la demande”, [...] an individual decision taken on the basis of an algorithmic treatment shall include an explicit mention by informing the person concerned. The rules defining this treatment and the main characteristics of its implementation are to be communicated by the administration to the person concerned if so requested [author translation]).

⁵³⁷ See further Edwards and Veale, ‘Enslaving the Algorithm’ (n 531).

⁵³⁸ Citron (n 535).

⁵³⁹ *ibid.*

⁵⁴⁰ *R (Moseley) v London Borough of Haringey* [2014] UKSC 56, 21. See generally Alistair Mills, ‘An Update on Consultation’ (2015) 20(3) Judicial Review 160 DOI: 10/gfsjqqs.

⁵⁴¹ GDPR, art 35(9).

⁵⁴² Binns, ‘Data protection impact assessments: A meta-regulatory approach’ (n 332) 28.

⁵⁴³ Citron (n 535).

2. The Law of Machine Learning?

consultation requirements above, is that data controllers are not required to publish the documents that result.⁵⁴⁴ This largely limits their (mandatory) functions as a i) trigger to consider data protection issues at an earlier stage of design and ii) a record which DPAs can hold organisations to account over. This might be slightly different for many public bodies, who are likely to find that their DPIAs are subject to FOI legislation. In some jurisdictions, it is also feasible that private bodies who have submitted their DPIAs to DPAs for the prior consultation procedure (mandatory where self-assessed risks are not self-assessed to be fully mitigated)⁵⁴⁵ might find those DPIAs to be FOI-able too.⁵⁴⁶ Yet this is likely to remain a specific minority (as few controllers are likely to wish to raise their heads above the parapet in favour of just self-assessing themselves more generously), and is especially unlikely to comprise processing seeking to exploit or manipulate individuals with machine learning that the controller may suspect could, in the eyes of an activist regulator, fall foul of the law.

2.4.3. Certification

A further novelty in the GDPR, and one promoting a move beyond individualistic provisions, are the new rules supporting the development of *certification systems*.

Conceptually, certification systems contain three elements: a *standard*, a *certificate*, and a *label*. A standard lists ‘specifications and/or criteria for the manufacture, use, and/or attributes of a product, process, or service’. Certification is the ‘process, often performed by a third party, of verifying that a product, process or service adheres to a given set of standards and/or criteria.’ Lastly, labelling is the ‘method of providing information on the attributes, often unobservable, for a product, process or service.’⁵⁴⁷

The certification provisions in the GDPR constrain the potential certification systems that might arise in certain, sometimes unhelpful, ways. Firstly, the standard specified *is* the GDPR and not beyond it in, for example, the way that Fair Trade certification attempts to go beyond national labour standards.⁵⁴⁸ Those being certified are controllers and processors, and cannot be individuals (in the sense of certifying them as *bona fide* consultants rather than snake-oil merchants—individuals can be controllers). Certification must be carried out by an independent third party (rather than

⁵⁴⁴ See later in this thesis for a discussion of publishing and DPIAs in section 3.1.5.3, p. 166.

⁵⁴⁵ GDPR, art 36. If such documents are subject to FOI requests (perhaps to different degrees in different jurisdictions), it is likely that this would further deter controllers from self-selecting for consultation, as they currently must.

⁵⁴⁶ As a side note, the Dutch have a great verb for making such requests—*wobben*, which means *to FOI request* (named after the law, the *Wet Openbaarheid van Bestuur* (WOB)).

⁵⁴⁷ Kira JM Matus, ‘Standardization, certification, and labeling: A background paper for the roundtable on sustainability workshop’ in Committee on Certification of Sustainable Products and Services (ed), *Certifiably Sustainable?* (National Academies Press 2010) DOI: 10/gfsjrf.

⁵⁴⁸ GDPR, art 42(1) (noting that certification is for the purpose of ‘demonstrating compliance’).

first party certification (self-regulation) or second party certification (certification by an organisation in an eg transactional relation)) accredited by a DPA or a designated national accreditation body in collaboration with a DPA.⁵⁴⁹ Labelling can then occur, and the EDPB has responsibility for collating certificates into a register and publishing them by any available means, presumably online.⁵⁵⁰

We have yet to see whether a certification market for machine learning and the GDPR takes off. All the provisions in this area depend on the action of supervisory authorities and of the EDPB, and the creation of well-funded or sensible schemes that result in significant uptake is far from assured. Indeed, the incentives for certification on the part of controllers and processors are at best mixed, as while certification is only concerned with whether a controller or processor *meets* the GDPR standard, certificates do not protect them against action resulting from the infringement of data protection law⁵⁵¹—which, as noted, is itself rife with grey areas and so limited assurance can be given of legality in the eyes of the CJEU.⁵⁵² However, the presence of a *bona fide* certificate may mitigate the gravity of regulatory penalty such as the extent of any fines levied.⁵⁵³

Certification might produce new avenues for transparency of broader machine learning development practices and processes to third parties rather than to data subjects.⁵⁵⁴ This is because the GDPR states that the controller or processor seeking certification must provide the accredited body with ‘all information and access to its processing activities which are necessary to conduct the certification procedure.’⁵⁵⁵ While this whole mechanism remains voluntary, it gives certification bodies that are able to achieve some sectoral recognition (and perhaps some expectation that firms *will* submit to certification) the power to act as an investigative force.

* * *

In sum, this chapter has outlined the prime candidates within the GDPR for governing machine learning systems that matter. Initially, the law appears to have a

⁵⁴⁹ The body in question must be the national body nominated under Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93 [2018] OJ L218/30.

⁵⁵⁰ GDPR, art 42(8). It is not wholly clear from the text whether it is the individual certificates that must be published or just the *available* mechanisms.

⁵⁵¹ GDPR, art 42(4).

⁵⁵² Damian Clifford and Jef Ausloos, ‘Technobabble and Technobullsh*t—what the hell is everyone on about?’ in *Presented at Gikii 2018, September 15 2017, Winchester, UK* (2017).

⁵⁵³ GDPR, art 83(2)(j).

⁵⁵⁴ See Tristan Henderson, ‘Does the GDPR Help or Hinder Fair Algorithmic Decision-Making?’ (LLM, University of Edinburgh 2017) DOI: 10/cx88.

⁵⁵⁵ GDPR, art 42(6).

2. The Law of Machine Learning?

lot of foresight in relation to the impact, or supposed concerns, around algorithmic systems—foresight that dates back some decades, in certain cases. On closer inspection however, there is an incongruity in both textual terms and when considered practically between the individual rights approach in the GDPR and the societal nature of the challenges machine learning systems appear to bring. This is compounded by the imagined form oversight takes, with issues raised by individual data subjects to regulators, rather than by civil society actors, who are comparatively disempowered in the enforcement process.

Going beyond these issues, it is possible to find ‘hooks’ in the law for a more collective flavour of oversight and action, in areas like purpose limitation, DPIAs and certification. Yet these do not come fully formed, and their efficacy in practice is far from assured.

Furthermore, and as I turn to now, machine learning does not only challenge the effective enforcement of data protection law, or highlight areas of rights and freedoms that are under-protected. I argue it has, by virtue of some of its technical characteristics, a destabilising effect on some of the underlying definitions and assumed notions that hold up the framework. In the next chapter, I tackle three main distinctions that data protection relies on both explicitly and implicitly, and how they are heavily blurred by machine learning in such a way that they risk eroding some of its protective or enforceable character.

3. Data Protection’s Lines, Blurred by Machine Learning

In chapter 2, *The Law of Machine Learning?*, I introduced the framework and varied provisions of the General Data Protection Regulation 2016 (GDPR), and explored how its provisions would apply, in more or less useful or protective ways, to machine learning technologies in the type of contexts and algorithmic systems described earlier in *Algorithmic War-Stories* (section 1.4). These included automated decision rights, access and explanation rights, and obligations on controllers that go beyond individualism.

Chapter 2 accepted machine learning technologies as more or less ‘playing nicely’ with European data protection law. It considered machine learning as an analytic product with impacts on individuals. This is a useful and popular lens, but I argue, does not depict the entire story. In this chapter, I argue there are unemphasised, novel tensions between data protection law and machine learning technologies, focussing, in a greater interdisciplinary light than chapter 2, on characteristics of the technology that I feel are neglected by the more conventional view on the law. These tensions between the technological practices and characteristics around machine learning are outlined in Figure 3.1. The analytical focus on these tensions remains legal, but necessarily delves into questions of practice and of the research around contemporary technological developments.

This unpacking proceeds in three parts.⁵⁵⁶ The first part, Line 1: People from Data, focusses on the *training* of machine learning models, and, in particular, how data controllers today are collecting and holding the datasets used for this training. I argue, drawing on vignettes of Apple’s *Siri* voice assistant and WiFi analysis undertaken by

⁵⁵⁶ The first two parts of this chapter have been published as Michael Veale, Reuben Binns and Jef Ausloos, ‘When data protection by design and data subject rights clash’ (2018) 8(2) *International Data Privacy Law* 105 DOI: 10/gdxthh (here, Line 1: People from Data) and Michael Veale, Reuben Binns and Lilian Edwards, ‘Algorithms That Remember: Model Inversion Attacks and Data Protection Law’ (2018) 376 *Phil. Trans. R. Soc. A* 20180083 DOI: 10/gfc63m (here, Line 2: Data from Models) respectively. They have been reworked, edited and updated for this thesis. A larger piece incorporating Line 3: Data from Sensitive Data has been presented at Privacy Law Scholars Conference Europe and at BILETA 2018 (Michael Veale and Lilian Edwards, ‘Better seen but not (over)heard? Automatic lipreading systems and privacy in public spaces’ [2018] Presented at PLSC EU 2018) but is unpublished.

3. Data Protection's Lines, Blurred by Machine Learning

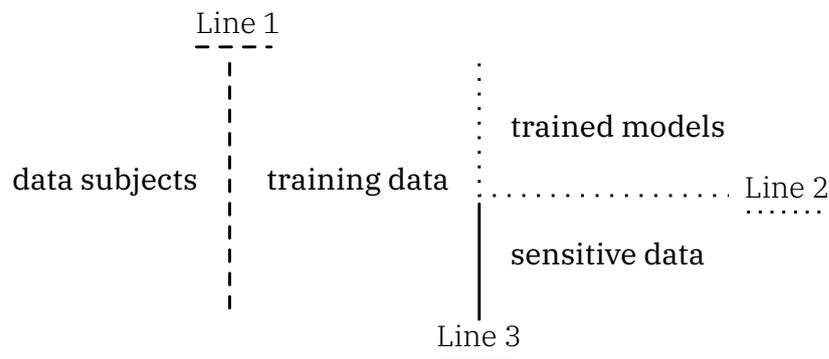


Figure 3.1.: Diagram of lines drawn by data protection and blurred by machine learning. Line 1 is covered in section 3.1, line 2 in section 3.2, line 3 in section 3.3.

Transport for London (TfL), that there is reason for concern in the particular ways that large datasets are amassed for the purposes of machine learning-based analytics. Data controllers are collecting and retaining sensitive data which is easy to re-identify for an adversary or ‘hacker’, but are, through varying methods, de-identifying it to the extent that the individuals represented in that dataset cannot re-identify themselves to the satisfaction of the data controller to exercise their data rights. This entails a trade-off between privacy-as-confidentiality, where even a small amount more effort spent on de-identification is justified, and what is arguably the essence of data protection, privacy-as-control, where it remains important that individuals can have some agency over data that relates to them, including to manage the risks it may pose to them.

The second part, Line 2: Data from Models, considers the legal status of machine learning in light of recent relevant cybersecurity literature. In this technical context, I argue that some machine learning models can, under existing law, be classified not just as an analytical product, but as *personal data in their own right*. This is due to their ‘leaky’ nature, which, for some models, allows attackers to extract an estimate of the original training data, or an understanding of who was present in the training set.

The third part, Line 3: Data from Sensitive Data, considers tensions around the distinction between *sensitive* (or special category) personal data, and ‘ordinary’ personal data in the GDPR. The potential of machine learning to transform such data was already the subject of the *Target* war story earlier,⁵⁵⁷ where loyalty card records were (allegedly) transformed into an estimate of the protected characteristic of pregnancy.

⁵⁵⁷ Section 1.4.3, p. 43.

In this section, I argue that machine learning facilitates more radical forms of data transformation that create large tensions across the framework. The worked example is of automated lipreading systems: software that transforms video into an estimated transcript of what was said—which resembles interception of communication, a power usually restricted to states (usually through cooperation with telecommunication providers), and something generally considered more severe than silent closed-circuit television (CCTV), at least in the UK setting where citizens are by-and-large acclimatised to the presence of ubiquitous cameras. The impact of this transformation is exacerbated by how data protection law already can be seen to fail to deal with text and transcripts, as the rich and contextual nature of their sensitivity arguably does not play nicely with the idea of more sensitive data being solely judged in the GDPR from ‘revealing’ certain *characteristics* about data subjects.

In all three parts, these tensions are outlined alongside their implications. In Line 1: People from Data, ways to better make trade-offs concerning machine learning datasets under the existing law are highlighted; in Line 2: Data from Models, the implications (and desirability) of extending new rights and obligations to data subjects is discussed; and in Line 3: Data from Sensitive Data, the methods data controllers might use to mitigate these tensions (successfully or unsuccessfully, or even to the detriment of data subjects wishing to avoid such surveillance) are discussed.

3.1. Line 1: People from Data

Nature of Line 1 The first line concerns the way data protection conceives of the utility of the separation of natural persons from the data about them. Separation of data from the people whom it is about is conceived in three core ways in the GDPR.

The first way, *pseudonymisation*, is a method by which identifiers are split from data records using internal ‘technical and organisational measures’ to ensure that additional data needed to attribute personal data to a specific data subject is ‘kept separately’.⁵⁵⁸ According to the recitals, it ‘can reduce the risks to data subjects’ and ‘help controllers and processors to meet their data protection obligation’.⁵⁵⁹ Such data can continue to be identified within a data controller (although there may be internal firewalls or organisational means preventing simple re-identification, such as giving the parts to different departments). A second level of separation can be seen for ‘[P]rocessing which does not require identification’, where for purposes which ‘do not or do

⁵⁵⁸ GDPR, art 4(5).

⁵⁵⁹ GDPR, recital 28.

3. Data Protection's Lines, Blurred by Machine Learning

no longer require the identification of a data subject by the controller',⁵⁶⁰ the controller is essentially partly relieved from fulfilling certain data subject rights, unless the data subject provides information which is 'enabling' of his or her identification.⁵⁶¹ Thirdly, the starkest (and difficult to achieve) separation is *anonymisation*, where data 'does not relate to an identified or identifiable natural person' or 'the data subject is not or no longer identifiable'.⁵⁶² The latter form of separation takes personal data outside of the GDPR entirely.⁵⁶³

These different levels seem increasingly protective of data subject interests as and if data controllers manage to move through them (a process moving from *deidentification* or the masking of direct identifiers to *anonymisation*, creating a negligible risk of reidentification⁵⁶⁴). This makes sense if, as is often the case, data are destined for release or further sharing after being subject to such safeguards.⁵⁶⁵ Yet in this section, this conception of the role of this separating line is challenged. Particularly in contexts involving processing, model training and machine learning, it is possible that deidentification can remove the ability for data subjects to be clearly or easily distinguished from other data for the purposes of exercising rights such as access, erasure or objection (which come with a high threshold demand for accuracy in order to stop leakage of others' data, or disruption of their services), but the same deidentification does not stop an individual being targeted, profiled or personalised too in the ways data protection (and the governance of machine learning in general) might seek to better control.

3.1.1. Introduction

Machine learning systems are generally built from extensive datasets, often about individuals.⁵⁶⁶ Those holding the datasets capable of fuelling the most advanced machine learning systems are considered to be among the most important players in the digital economy. This is of societal concern for a number of core reasons.

⁵⁶⁰ GDPR, art 11.

⁵⁶¹ GDPR, art 11(2).

⁵⁶² GDPR, recital 26.

⁵⁶³ It is however worth noting that if the original dataset still exists, the anonymised dataset is arguably still personal according to the reidentification standards established by the CJEU (and outlined further below, see section 3.2.4.) This remains a tricky tension, even described as a paradox by some. See Mark Elliot, Elaine Mackey, Kieron O'Hara and Caroline Tudor, *The Anonymisation Decision-Making Framework* (UKAN 2016) 10.

⁵⁶⁴ *ibid* 15–16.

⁵⁶⁵ See *ibid* 13.

⁵⁶⁶ This data is almost always collected from the 'real world'; in rare cases, such as in simulations of computer games such as Google Deepmind's *Alpha Go*, it is largely collected through simulation (eg a computer playing another computer). While these techniques may show promise for consequential societal applications in the future, they currently are not heavily used, if at all, outside of research contexts, so are not discussed in this thesis.

Firstly, there are market and competition aspects. Regulators and international bodies are increasingly growing interested in competition issues around data,⁵⁶⁷ as are researchers looking at data accumulation by actors such as data brokers online.⁵⁶⁸ Whether such collection and retention of data effectively precludes effective market functioning is a hot topic of discussion in national and international governments.

Secondly, there are security aspects. Naturally, a trove of data so extensive is an appealing target for attack and acquisition. Insofar as it contains aspects about individuals they cannot effectively change to mitigate any impact—such as their biometrics, or their communications history—the impact of a breach against the largest actors could, in short, change the world. Such an earthquake is commonly discussed in considerations of post-quantum encryption, where some argue that were successful quantum computers to be invented, much data that is currently encrypted using standard measures would be vulnerable to being revealed. A relevant war-story closer to today is the ‘data breach’ at Facebook that led to the release of user data used to create micro-targeting tools for political contexts, which some have argued influenced democratic processes around the world.⁵⁶⁹ Not only did this illustrate the loss of control that data subjects experienced over their data, which was effectively and unbeknownst to them being piped out of the platform via application programming interfaces (APIs) research institutions and companies were granted access to by users’ friends—it also illustrated how such accumulated data might be weaponised using machine learning technologies, and purposed both against the very users it was stolen from and against society more generally. The collapse of control at Facebook (they have been fined the maximum permissible fine of 500,000 GBP by the ICO⁵⁷⁰), and the relevant collapse of trust in its ability to secure data by its user-base,⁵⁷¹ has illustrated the risks that the

⁵⁶⁷ Antonio Capobiano, Pedro Gonzaga and Anita Nyeső, *DAF/COMP(2017)4: Algorithms and collusion - Background note by the Secretariat* (Organisation for Economic Co-operation and Development (OECD) 2017) ([https://one.oecd.org/document/DAF/COMP\(2017\)4/en/pdf](https://one.oecd.org/document/DAF/COMP(2017)4/en/pdf)); Thiemann, Gonzaga and Stucke (n 285); European Data Protection Supervisor, *Privacy and Competitiveness in the Age of Big Data: The Interplay between Data Protection, Competition Law and Consumer Protection in the Digital Economy* (n 285); *Big Data und Wettbewerb* (n 285); *Big Data and Competition* (n 285).

⁵⁶⁸ See eg Reuben Binns, Jun Zhao, Max Van Kleek and Nigel Shadbolt, ‘Measuring third party tracker power across web and mobile’ (2018) 18(4) *ACM Transactions on Internet Technology (TOIT)* 52; Elettra Bietti and Reuben Binns, ‘Acquisitions in the Third Party Tracking Industry: Competition and Data Protection Aspects’ [2018] Preprint available on SSRN (<https://papers.ssrn.com/abstract=3269473>); Reuben Binns, Ulrik Lyngs, Max Van Kleek, Jun Zhao, Timothy Libert and Nigel Shadbolt, ‘Third Party Tracking in the Mobile Ecosystem’ in *Proceedings of the 10th ACM Conference on Web Science* (ACM 2018) DOI: 10/cwdk; Maurice E Stucke and Allen P Grunes, ‘No Mistake About It: The Important Role of Anti-trust in the Era of Big Data’ [2015] U Tennessee Legal Studies Research Paper.

⁵⁶⁹ See section 1.4.5, p. 47. Note that some argue the opposite, that such practices were of little consequence. See eg David A Graham, ‘Not Even Cambridge Analytica Believed its Hype’ (*The Atlantic*, 20th March 2018) (<https://perma.cc/7G5B-2PPF>). We will likely never know for certain.

⁵⁷⁰ See the penalty notice at <https://perma.cc/TM4F-DGEF>.

⁵⁷¹ Hannah Kuchler, ‘Zuckerberg failed to fix Facebook users’ privacy concerns’ (*Financial Times*, 17th April 2018) (<https://perma.cc/34EM-DQF7>).

3. Data Protection's Lines, Blurred by Machine Learning

monopoly data holdings of such technology giants might pose.

Thirdly, there are privacy and information asymmetry aspects connected to the potential for inference such data mountains afford. Data from users' activities, such as their click-patterns, telemetry from their phone such as gyrometer, compass and settings, location traces from their use of navigational tools, comments on their friends' online content and more all might seem relatively straightforward in their consequences when given to a data controller in isolation. Conventional analysis this data could be subject to might be reasonably foreseen by a data subject, but machine learning systems have demonstrated quite clearly that in combination with huge datasets with varied information about thousands, if not millions, of data subjects, the analysis that this could be subject to is hardly predictable. For controllers with such data mountains, the appropriate question to ask almost now seems less like what might be able to be predicted with such data, but what data controllers think they cannot predict with it.⁵⁷²

Data protection law has within it provisions to address all these concerns, at least to some degree. Several actors, including the Commission, connect the *right to data portability*⁵⁷³ to market dominance, suggesting that it might help achieve certain aims of competition law.⁵⁷⁴ Security provisions in data protection, included heightened obligations around data breach notification (which lacked a pan-European harmonised approach before the GDPR and did not feature in the DPD) as well as the overarching security principle and penalties for non-compliance all incentivise better security. Purpose limitation is supposed to prevent creep from one use of data to another and ensure processing stays broadly within data subjects' expectations.

Further to both security and privacy in particular, data subjects are supposed to be able to vote with their feet. Their rights of erasure, objection, portability and the like are in part to allow them to remove their data or limit its use for controllers they do not or no longer trust, both for their own autonomy, and arguably as a threatening mechanism for controllers to not just stay in line with the law, but also keep individuals more deeply on board.

This section however argues that the mechanisms by which data mountains are built and maintained in practice threaten such control rights. Some data controllers interested in amassing large datasets, often for machine learning or personalisation services, appear to now do so in a way where, despite such data remaining personal under data protection law, it is difficult—they argue impossible—for data subjects to ex-

⁵⁷² Whether these predictions 'work' is a different question entirely.

⁵⁷³ GDPR, art 20.

⁵⁷⁴ See generally Helena Ursic, 'Unfolding the New-Born Right to Data Portability: Four Gateways to Data Subject Control' (2018) 15(1) SCRIPTed 42 DOI: 10/gfc7c9.

ercise their rights over. This is achieved by effectively deleting or otherwise masking the *explicit identifier* in these datasets, such as name, email, or identification number, and leaving individuals in a situation where they are unable to pick themselves out of the dataset. This initially seems like a useful activity for data controllers to engage in: in line with the principles of data minimisation⁵⁷⁵ and purpose limitation.⁵⁷⁶ It also could be argued as in line with the notion of privacy-by-design now explicitly part of European data protection law. However, on closer inspection, the chosen term and framing for ‘privacy-by-design’ in data protection is *not* about keeping data as confidential as possible—the traditional aim of privacy-enhancing technologies (PETs)—but a more holistic notion of control in relation to the *fundamental right of data protection*.

What is the challenge here? Data protection law is, in some ways, a naïve framework. Its architects largely imagined its provisions as if they were all mutually complementary to the Regulation’s goals, all of the time, rather than as presenting value-laden and often difficult trade-offs to designers. Consequently, while the law is littered with different flavours of transparency mechanism,⁵⁷⁷ these mechanisms do not primarily help in making the way these trade-offs are navigating explicit.

The result of this, as will be shown in two case studies, is a situation where a large amount of data is stored by a data controller in a position of market or other power largely for the purpose of machine learning, where such data remains relatively trivial to re-identify to a natural person (in the sense of technical feasibility) and to reveal sensitive insights from; but where these natural persons have been effectively stripped of their rights over such data, losing the ability to manage the risk it poses themselves. In some cases, the data controller surprisingly *retains* the technical ability to target the user with individualised personalisation, creating a large and surprising power asymmetry that is enabled by PETs.

3.1.2. Background

Data protection law has historically faced significant enforcement challenges. DPAs have classically been underfunded and outgunned, possessing limited ability to scrutinise the on-the-ground practices of data controllers and restricted capacity to meaningfully act when transgressions are suspected. In response to these governance challenges, concerned communities have advocated a range of technological approaches

⁵⁷⁵ GDPR, art 5(1)(c), stating that data must be ‘adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed’.

⁵⁷⁶ GDPR, art 5(1)(b), stating that data must be ‘collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes’.

⁵⁷⁷ See eg section 2.2.2.

3. Data Protection's Lines, Blurred by Machine Learning

that allow effective but non-invasive use of data, or 'DIY' protections which data subjects can adopt unilaterally.⁵⁷⁸

These approaches—privacy-enhancing technologies (PETs)—are commonly discussed in regulatory circles within the context of privacy by design (PbD). PbD emphasises that issues of privacy should be considered from the start and throughout the design process through creative social and technical means. Most point to its intellectual home in a report undertaken by the Dutch DPA and the TNO, with support of the Information and Privacy Commissioner for Ontario,⁵⁷⁹ although its heritage can be traced further back to the considerations given to “technical and organisational measures” in the DPD⁵⁸⁰ and in the national and regional laws that preceded it⁵⁸¹. The term PbD entered use around 2000, with the Workshop on Freedom and Privacy by Design at the Computers, Freedom and Privacy 2000 conference in Toronto⁵⁸² and a variety of papers made use of the term around that time.⁵⁸³ As laid out by the Information and Privacy Commissioner for Ontario from 1998–2014, Ann Cavoukian, PbD is not simply a set of organisational and technical measures to prevent information disclosure, but maps more broadly onto a wider idea of privacy as represented by the Fair Information Practices (FIPs) and even extends beyond them, aiming at a ‘significant ‘raising’ of the bar in the area of privacy protection’.⁵⁸⁴

While recommendations of PbD by regulators have significant history,⁵⁸⁵ the concept has only recently made it onto the statute books in Europe as part of the GDPR. In doing so, it underwent a shrewd transformation into data protection by design (DPbD). This metamorphosis, which some scholars have commented on as wise,⁵⁸⁶ makes it clear that the aim is to ensure privacy as enshrined in data protection rights and principles, rather than the flexible, multi-layered and tricky-to-pin-down concept of privacy in general.⁵⁸⁷ While the European Commission has historically referred to

⁵⁷⁸ Claudia Diaz, Omer Tene and Seda Gürses, ‘Hero or villain: The data controller in privacy law and technologies’ (2013) 74 Ohio St. LJ 923.

⁵⁷⁹ Information and Privacy Commissioner of Ontario, Canada and Registratiekamer, the Netherlands, *Privacy-Enhancing Technologies: The Path to Anonymity* (Information and Privacy Commissioner and Registratiekamer 1995) (<http://govdocs.ourontario.ca/node/14782>).

⁵⁸⁰ DPD, recital 46; DPD, ¶ 17.

⁵⁸¹ See generally González Fuster (n 295).

⁵⁸² See www.cfp2000.org/program/full-program.html.

⁵⁸³ See eg Julie E Cohen, ‘Examined lives: Informational privacy and the subject as object’ (2000) 52 Stanford Law Rev. 1373; Marc Langheinrich, ‘Privacy by Design — Principles of Privacy-Aware Ubiquitous Systems’ in Gregory D Abowd, Barry Brumitt and Steven Shafer (eds), *Proceedings of Ubicomp 2001* (Springer 2001) vol 2201.

⁵⁸⁴ Ann Cavoukian, *Privacy by Design: The 7 Foundational Principles* (Information and Privacy Commissioner for Ontario 2010) 1.

⁵⁸⁵ See eg Information and Privacy Commissioner of Ontario, Canada and Registratiekamer, the Netherlands (n 579).

⁵⁸⁶ Mireille Hildebrandt and Laura Tielemans, ‘Data protection by design and technology neutral law’ (2013) 29(5) Comput. Law & Secur. Rev. 509.

⁵⁸⁷ See generally Kieron O’Hara, ‘The Seven Veils of Privacy’ (2016) 20(2) IEEE Internet Computing 86 DOI:

the two concepts synonymously,⁵⁸⁸ the focus on DPbD alone provides scope for further clarity. Lee Bygrave summarises that DPbD requirements, as now enshrined in both the GDPR⁵⁸⁹ and the Law Enforcement Directive,⁵⁹⁰ impose a ‘qualified duty on controllers to put in place technical and organisational measures that are designed to implement data protection principles effectively and to integrate necessary safeguards into the processing of personal data so that such processing will meet the Regulation’s requirements and otherwise ensure protection of data subjects’ rights’.⁵⁹¹ In the GDPR, the DPbD⁵⁹² provisions read:

1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.
2. The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual’s intervention to an indefinite number of natural persons.
3. An approved certification mechanism pursuant to Article 42 may be used as an element to demonstrate compliance with the requirements set out in paragraphs 1 and 2 of this Article.

What are these principles that technical and organisational measures should take

10/gfbzpw.

⁵⁸⁸ See eg European Commission, *A Digital Agenda for Europe (COM(2010)245 final)* (2010).

⁵⁸⁹ GDPR, art 25.

⁵⁹⁰ Law Enforcement Directive, art 20.

⁵⁹¹ Lee A Bygrave, ‘Data Protection by Design and by Default : Deciphering the EU’s Legislative Requirements’ (2017) 1(2) *Oslo Law Review* 105, 114.

⁵⁹² GDPR, art 25.

3. Data Protection's Lines, Blurred by Machine Learning

aim at? They are found primarily in the data protection principles: lawful, fair and transparent processing; purpose limitation; data minimisation; accuracy; storage limitation; and integrity and confidentiality,⁵⁹³ plus the overarching principle of 'accountability',⁵⁹⁴ laying the burden of proof on the controller to prove compliance with the six 'main' principles.

Yet in contrast to these wide-ranging principles, within which reside the rights and obligations the legislation details, the PETs literature takes relatively single-minded aim at information disclosure. It focusses in particular on guarantees rooted in either information theory or the computational 'hardness' of the resultant re-identification or disclosure problem.⁵⁹⁵ Despite attempts in related literature on complementary approaches to coin terms such as 'transparency-enhancing technologies' and 'profile transparency by design', the PET paradigm has dominated the 'by design' discussion in data protection contexts.⁵⁹⁶ Unlike the data protection paradigm, which has increasingly shifted to placing accountability obligations upon data controllers in an effort to make them trusted custodians of personal data, the PET paradigm departs from a 'diametrically opposed perception', not of the data controller as a trusted third party, but as an adversary.⁵⁹⁷ In a similar vein, recent taxonomies of privacy-enhancing technologies claiming to be 'comprehensive' consider privacy primarily in terms of disclosure risks present at different levels, rather than in terms of the multi-faceted nature of privacy espoused by Cavoukian and the European Commission.⁵⁹⁸

This notion of privacy-as-confidentiality sits at least apart from, and potentially at tension with, the notion of privacy-as-control as espoused by the FIPs and the GDPR.⁵⁹⁹ As the A29WP notes, PbD incorporates rights such as erasure, noting that "functionality should be included facilitating the data subjects' right to revoke consent, with subsequent data erasure in all servers involved (including proxies and mirroring)." They note that in addition to data confidentiality, 'controllability', 'transpar-

⁵⁹³ GDPR, art 5(1).

⁵⁹⁴ GDPR, art 5(2).

⁵⁹⁵ See generally Casey Devet and Ian Goldberg, 'The Best of Both Worlds: Combining Information-Theoretic and Computational PIR for Communication Efficiency' in *Privacy Enhancing Technologies* (Springer 2014).

⁵⁹⁶ See eg Mireille Hildebrandt, 'Profile Transparency by Design?: Re-Enabling Double Contingency' in Mireille Hildebrandt and Katja de Vries (eds), *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology* (Routledge 2013); Milena Janic, Jan Pieter Wijbenga and Thijs Veugen, 'Transparency Enhancing Tools (TETs): An Overview' in *Third Workshop on Socio-Technical Aspects in Security and Trust* (2013) DOI: 10/cwmv.

⁵⁹⁷ Diaz, Tene and Gürses (n 578).

⁵⁹⁸ Johannes Heurix, Peter Zimmermann, Thomas Neubauer and Stefan Fenz, 'A taxonomy for privacy enhancing technologies' (2015) 53 *Comput. Secur.* 1 DOI: 10/f74pgf, 1. Note that not all conceptions of privacy engineering share these assumptions: cf Marit Hansen, Meiko Jensen and Martin Rost, 'Protection Goals for Privacy Engineering' in *2015 IEEE Security and Privacy Workshops* (IEEE 2015) DOI: 10/cwmw.

⁵⁹⁹ Seda Gürses, 'Can You Engineer Privacy?' (2014) 57(8) *Commun. ACM* 20 DOI: 10/gdxwh5, 20.

ency’, ‘data minimisation’ and ‘user-friendly systems’ should be considered under the PbD umbrella.⁶⁰⁰

Despite these clarifications by regulators, the re-naming of the term to emphasise its focus, and the commentary in the literature on the wide array of protection goals that privacy engineering should have,⁶⁰¹ PbD in practice is often a narrower affair. Where data are of high dimensionality (where they have many distinct variables), many PbD approaches aimed at the ‘unlinkability’ of data⁶⁰² will inevitably fail to prevent information disclosure where faced with a capable adversary. This does not mean that PETs cannot be used to minimise or reduce risk in this way, but I argue that this minimisation comes at a cost. That cost can be, as demonstrated below with case studies, the effective ability to wield data protection rights—the *intervenability* promoted by the fundamental right to data protection⁶⁰³—over such data. The important rights of access and portability,⁶⁰⁴ erasure,⁶⁰⁵ and the right to object to processing⁶⁰⁶ suffer in particular as a result.

There is a danger that data controllers implement privacy design strategies⁶⁰⁷ that leave them with data that is difficult for them to re-identify, but far from trivial for an adversary to do so, given that adversaries likely have a high tolerance for inaccuracy and access to many additional, possibly illegal, databases to triangulate individuals with. The situation is worsened by the fact that a data controller may have relatively little technical re-identification capacity, whilst also having a very low tolerance for inaccuracy when it comes to their provision of core data protection rights, such as access or erasure. Indeed, to erroneously provide a data subject sensitive personal data of another in response to a subject access request would usually be in breach of the very GDPR that the controller would be seeking to comply with.

These controllers, some of which will now be illustrated below, have bound their hands in a very particular way. Their actions have reduced their own data protection obligations and shifted a risk onto the data subject, who has been stripped of her ability to manage the risk herself. When the data subject concerned loses trust in a previously

⁶⁰⁰ Article 29 Data Protection Working Party, *The Future of Privacy: Joint contribution to the Consultation of the European Commission on the legal framework for the fundamental right to protection of personal data (02356/09/EN WP 168)* (2009) 14.

⁶⁰¹ Information and Privacy Commissioner of Ontario, Canada and Registratiekamer, the Netherlands (n 579).

⁶⁰² Andreas Pfitzmann and Marit Köhntopp, ‘Anonymity, Unobservability, and Pseudonymity – A Proposal for Terminology’ in *Designing Privacy Enhancing Technologies* (Springer 2001).

⁶⁰³ See section 2.1.

⁶⁰⁴ GDPR, art 15; GDPR, art 20.

⁶⁰⁵ GDPR, art 17.

⁶⁰⁶ GDPR, art 21.

⁶⁰⁷ See further and generally Jaap-Henk Hoepman, ‘Privacy Design Strategies’ in *ICT Systems Security and Privacy Protection* (Springer 2014).

trusted controller; there is nothing she can do but wait for a breach and hope that her record is unable to be effectively triangulated.

I do not intend to suggest that this is a deliberate tactic by the data controllers in our case studies (even though it might be an effective one). However, it *does not need to be deliberate to be problematic*. Trade-offs are a natural part of all complex decision-making, and the need to make them clearly rather than implicitly is a core component of good decision-making in value-laden contexts.⁶⁰⁸ Where there are very few organisational or technical measures supporting data protection deployed, DPbD is likely to benefit everyone. But where basic safeguards are already in place, satisfying everyone and their varying privacy preferences;⁶⁰⁹ may become more difficult, as 'privacy' is no longer a case of Pareto-improvement (under which it can masquerade as a unified concept), but can require choosing a certain approach (eg confidentiality) to the detriment of another (eg control). Thinking in terms of data protection rights and obligations as done here can make this challenge clearer: achieving one makes it more difficult, or even impossible, to achieve others. Not engaging with these trade-offs does not make them disappear, it simply means they have been determined in an arbitrary fashion. Here, the vignettes presented do indicate that certain controllers pursue an interpretation of these provisions, deliberately or not, which is unfavourable to the effective exercise of data subject rights.

Deliberate or not, these implicit trade-offs are not even contemplated by preemptive provisions in data protection law, such as DPIAs. All is not lost, however. There are indeed grounds in the GDPR to support more consideration and transparency regarding the way these trade-offs are determined and communicated—and it is important that these are identified and put to use—but it requires new readings of many of the relevant obligations which this section aims to provide. Firstly however, I turn to two real-world case studies—vignettes of vanished rights—to explore this concern in context.

3.1.3. Vignette of Vanished Rights 1: TfL Wi-Fi

Between 21 November and 19 December 2016, Transport for London (TfL), the public transit agency for the UK's capital, ran an in-house trial using the Wi-Fi networks installed at 54 of the stations they manage. They collected more than 500 million connection requests from devices passively transmitting their media access control (MAC)

⁶⁰⁸ The ubiquity of trade-offs and the importance of making them explicitly is a core component of public policy education. See eg Eugene Bardach, *A practical guide for policy analysis* (SAGE 2012).

⁶⁰⁹ On varying privacy preferences, see eg Mark S Ackerman, Lorrie Faith Cranor and Joseph Reagle, 'Privacy in e-commerce: examining user scenarios and privacy preferences' (1999) DOI: 10/bdgc4r.

addresses, with the aim of improving 1) customer information for journey planning and congestion; 2) management of events and disruption; 3) timetable planning and station upgrades; 4) retail unit and advertising positioning.⁶¹⁰

Transport for London, unlike many undertaking Wi-Fi analytics,⁶¹¹ were aware of legal obligations in this area, data protection in particular. TfL undertook a DPIA and met with the UK's DPA, the ICO.⁶¹² They cite the ICO's WiFi Analytics Guidance⁶¹³ in their use of salting MAC addresses to make re-identification on the basis of device hardware data highly challenging for an attacker. In consultation with the ICO, users were informed using a 'layered approach', which included a press release picked up by the media, a news story on 21 November in the Metro (a free morning newspaper widely distributed and read on London transport), a linked website⁶¹⁴ adapted throughout the trial on the basis of feedback with users, 300 large posters on platforms and at station entrances, through social media and through briefings packs issued to station staff and stakeholder organisations.⁶¹⁵

As location data is high dimensional, it is highly likely to be unique and easy to re-identify. A now classic study showed that only four spatiotemporal points are needed to single out the vast majority of individuals in a dataset, even where records are rendered significantly coarser (something that often heavily diminishes the data's utility).⁶¹⁶ Unsurprisingly, TfL is therefore uncomfortable with releasing the dataset, refusing it on privacy grounds when requested under FOI law. When asked, they note (in my view, correctly) that:

Although the MAC address data has been pseudonymised [...] given the possibility that the pseudonymised data could, if it was matched against

⁶¹⁰ Transport for London, *Insights from Wi-Fi Data: Proposed Pilot* (TfL (Released under the Freedom of Information Act 2000) 2016) (<https://perma.cc/NSS3-7RW5>) 6–8.

⁶¹¹ See eg College bescherming persoonsgegevens, *Wifi-tracking van mobiele apparaten in en rond winkels door Bluetrace (Rapport z2014-00944)* (Autoriteit Persoonsgegevens (Dutch Data Protection Authority) October 2015) (<https://perma.cc/6BKJ-JGPY>), where the Dutch DPA examined WiFi analytics company BlueTrace and found them non-compliant with data protection law, with little chance of changing their business model to become so—they eventually closed rather than finding a means to adequately comply. See further Vasilios Mavroudis and Michael Veale, 'Eavesdropping Whilst You're Shopping: Balancing Personalisation and Privacy in Connected Retail Spaces' in *Proceedings of the 2018 PETRAS/IoTUK/IET Living in the IoT Conference* (IET 2018) DOI: 10/gffng2.

⁶¹² Transport for London, *Review of the TfL WiFi Pilot* (TfL 2017) 22

⁶¹³ Information Commissioner's Office, *Wi-fi location analytics* (ICO 2016) (<https://ico.org.uk/media/for-organisations/documents/1560691/wi-fi-location-analytics-guidance.pdf>).

⁶¹⁴ <http://tfl.gov.uk/privacy>.

⁶¹⁵ See Transport for London, *Review of the TfL WiFi Pilot* (n 612); Transport for London, *TfL WiFi Analytics Briefing Pack* (November 2016) (<https://perma.cc/7PHN-WBGH>). Note that some civil society organisations felt that the posters displayed in and around stations were insufficiently clear about how to opt-out. See Ed Johnson-Williams, 'TfL needs to give passengers the full picture on WiFi collection scheme' [2016] Open Rights Group Blog (<https://perma.cc/8YEA-BV8D>).

⁶¹⁶ Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen and Vincent D Blondel, 'Unique in the Crowd: The privacy bounds of human mobility' (2013) 3 *Sci. Rep.* 1376 DOI: 10/msd.

3. Data Protection's Lines, Blurred by Machine Learning

other data sets, in certain circumstances enable the identification of an individual, it is personal data. The likelihood of this identification of an individual occurring would be increased by a disclosure of the data into the public domain, which would increase the range of other data sets against which it could be matched.⁶¹⁷

Some concerns have been raised over the nature of the 'salt' added to the MAC address or other identifier to generate the string to be hashed.⁶¹⁸ Whilst the ICO recommends that a salt be changed after 'a short period of time'⁶¹⁹ and the Article 29 Working Party recommends that a unique device identifier should only be stored 'for a maximum period of 24 hours for operational purposes',⁶²⁰ it appears that TfL used a constant salt, generated by once typing letters at random on the keyboard with averted eyes.⁶²¹ Such an approach creates two risks. First, anyone who knew or discovered this salt could reverse engineer the process. Second, and arguably more probably, a constant salt links devices across days, making attacks not aimed at cryptography but based on external sources of data, such as knowing where someone was at four particular times in a week, more feasible.

One approach would seek to make extra efforts to de-identify the held data. The main way to make data more difficult to re-identify would be to give records more frequently-changing, difficult-to-reverse hashed identifiers. But this would likely be unacceptable for some data controllers, as it makes the purpose of the analysis they seek to undertake difficult to fulfil, and so data subjects might suspect that data controllers would wish to transform the data in this way. For example, it would preclude the use of analysis to understand longitudinal patterns in data, restricting them only to what can be learned in a snapshot of time. This is far from the logic of the A/B testing style trials favoured in both industry and policy circles right now.⁶²² Furthermore,

⁶¹⁷ Natasha Lomas, 'How "anonymous" wifi data can still be a privacy risk' [2017] TechCrunch (<https://perma.cc/Y63T-MAC8>).

⁶¹⁸ A hash function is a one-way transformation of data. For example, the md5 (a *message digest* algorithm) hash of 'iheartdataprotection' is '374d67ace049664f8837250bab7010ed'. A salt is a string added to data before it is hashed. For example, to add the salt '1' would result in 'iheartdataprotection1', which has a different md5 hash ('d6790618285a4f41c79aba2eb9bced3e'). There should be no reversible mathematical link between those two outputs; the only way to reverse engineer is through 'brute force'. Yet as someone could (and people do) make extremely large tables of all possible MAC addresses and their resultant hashes, salts are crucial to avoid reversal of the hash process.

⁶¹⁹ Information Commissioner's Office, *Wi-fi location analytics* (n 613) 6.

⁶²⁰ Article 29 Data Protection Working Party, *Opinion 13/2011 on Geolocation services on smart mobile devices' (WP 185) (2011) 19.*

⁶²¹ Lukasz Olejnik, 'Privacy of London Tube Wifi Tracking' (*Security, Privacy & Tech Inquiries [blog]*) (<https://blog.lukaszolejnik.com/privacy-of-london-tube-wifi-tracking/>); Lomas (n 617).

⁶²² While snapshot analytics might help an organisation like TfL better understand overcrowding and crowd management, for example, it would not, for example, allow them to easily understand something such as whether individuals that often run down escalators that subsequently stop by certain posters telling them not to indeed change their behaviour in the future.

and to the point of this thesis, prediction over time is precisely what machine learning is used mostly for, both for personalisation of individuals and building up profiles of stereotypical behaviour over and across time. Snapshot data precludes a lot of machine learning-powered analysis, and as such it is very difficult for data controllers to de-identify or ‘sanitise’ datasets beyond a certain point.

Yet another approach sits on the side of the data subject, rather than the controller. More specifically, it sits with capabilities and behaviours of the hardware used. Much of data protection law aims to build trust in data controllers as responsible stewards of sensitive information. Yet proponents of personal PETs take what some may consider as a contrasting, comparatively dismal view of the world—a gloomy planet where every other actor is a potential adversary that wants to do harm to them with their data—and as such seek to adopt technical practices in order to minimise the information that any third party can learn about them. These practices are increasingly popular with some software and hardware producers. Apple’s portable devices include MAC address randomisation, which seeks to foil third parties working to build a longitudinal record of a particular device’s network scanning activity. Some Android devices utilise this, although many manufacturers, such as Samsung, do not support or practice it.⁶²³ This has a similar, although not identical,⁶²⁴ effect to regularly changing the salt, and serves to make persistent tracking harder.

Yet even with these approaches enabled, researchers are consistently finding ways, both statistical and based on technical implementation or other features of smartphones, to link individuals across contexts.⁶²⁵ As TfL recognise, despite protections placed at either the controller side or the device side, such data is not safe from re-identification attacks.

Given this risk of reidentification, particularly from adversaries were data to leak, does a data subject not have a right to understand the data that is being collected about them, and utilise their rights, such as the right to object to processing, or the right to erase data relating to them? It is not difficult to imagine a situation where a previously trusted data controller now loses trust, either to be a well-intentioned custodian of data, or to be capable of keeping it confidential with high certainty.⁶²⁶ While a data

⁶²³ See Jeremy Martin and others, ‘A Study of MAC Address Randomization in Mobile Devices and When it Fails’ (2017) 2017(4) Proceedings on Privacy Enhancing Technologies (PoPETs) 268 DOI: 10/cwm4.

⁶²⁴ In particular, MAC randomisation does not prevent attackers recovering the several MAC addresses from unsalted or poorly salted hashes through brute force.

⁶²⁵ See eg Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S Cardoso and Frank Piessens, ‘Why MAC Address Randomization is Not Enough: An Analysis of Wi-Fi Network Discovery Mechanisms’ in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security (ASIACCS’16)* (ACM 2016).

⁶²⁶ It could be argued that as TfL already have potentially re-identifiable gate-to-gate data on users travel behaviour collected by their Oyster smart ticketing system, such erasure would often not do much to reduce data on their behaviour from data controllers, unless individuals relied on higher-cost disposable

3. Data Protection's Lines, Blurred by Machine Learning

subject may well wish to do so, these protections, whilst not fully mitigating any risk, do effectively remove the ability of data controllers to provide the full range of data protection rights usually afforded to data subjects. Indeed, TfL note:

The salt is not known by any individual and was destroyed on the day the data collection ended. Therefore, we consider the data to be anonymous and are unable to identify any specific device. As we cannot process known MAC addresses in the same manner as we did in the pilot, we are unable to complete any Subject Access Request for the data we collected.⁶²⁷

Were TfL to attempt this, they would find that in the cases of some hardware, the difficulty would be compounded by the device MAC randomisation practices described above. In particular, while devices can be identified with acceptable levels of accuracy for an attacker,⁶²⁸ the levels of identification achieved would be insufficient for providing guaranteed and comprehensive erasure, or accurate access (including avoiding divulging information about others).⁶²⁹ This reduces the protection afforded by law to the security provisions in the GDPR, as well as the trust in the controller to adhere to the principle of purpose limitation, giving the data subject little-to-no control over the data observed about them after the fact.

As mentioned, beyond subject access requests, another provision in the GDPR relates to the right to object to processing.⁶³⁰ Where the legitimate or public interest grounds are relied upon, data subjects should be 'entitled to object to the processing of any personal data' unless the controller can 'demonstrate that its compelling legitimate interest overrides the interests or the fundamental rights and freedoms of the data subjects'.⁶³¹ This manifests as an 'opt-out' provision, recommended in relation to 'big data' analytics grounded in legitimate interests by both the ICO and the EDPS.⁶³² Opt-outs from WiFi analytics in particular feature in the ICO's guidance on the matter,⁶³³ although whether they are mandatory under European law is unclear.⁶³⁴ This

paper tickets. However, WiFi analytics would provide re-identification capacity above this, and could even reveal additional information, such as adverts looked at, or which individuals were travelling in proximity to each other.

⁶²⁷ Transport for London, *Review of the TfL WiFi Pilot* (n 612) 22.

⁶²⁸ One study found identification success ranged from around 20-50% in the presence of MAC randomisation, becoming more difficult with more individuals present, and increased time of tracking demanded. See Vanhoef, Matte, Cunche, Cardoso and Piessens (n 625).

⁶²⁹ On the risks of subject access requests creating privacy breaches, see Andrew Cormack, 'Is the Subject Access Right Now Too Great a Threat to Privacy' (2016) 2 Eur. Data Prot. L. Rev. 15.

⁶³⁰ GDPR, art 21.

⁶³¹ GDPR, recital 69.

⁶³² Information Commissioner's Office, *Big data, artificial intelligence, machine learning and data protection* (n 128) para 69; European Data Protection Supervisor, *Meeting the challenges of big data, Opinion 7/2015* (EDPS 2015).

⁶³³ Information Commissioner's Office, *Wi-fi location analytics* (n 613)

⁶³⁴ Berber F E Bosch and Nico A N M van Eijk, 'Wifi-tracking in de winkel(straat): inbreuk op de privacy?'

may yet change in the proposed updated ePrivacy regulation (at the time of writing in trialogue negotiation), which has been amended to require opt-outs when WiFi analytics have been used.⁶³⁵ The Dutch DPA stopped short of mandating WiFi tracking firm *Bluetrace* to be required to offer opt-outs, instead settling for the company to undertake research into their technical feasibility, and pointing them towards opt-out registers being developed by Dutch civil society organisations.⁶³⁶ To the author's knowledge, the company never did so, having instead opted to cease WiFi analytics entirely, its business model being incompatible with the requirements of the regulator.⁶³⁷

In addition, the 'Mobile Location Analytics Code of Conduct' proposed by the Future of Privacy Forum (FPF) has opting-out as one of its principles, noting that the option should be available on the website of an operator.⁶³⁸ Indeed, FPF themselves run an opt-out service which partners with some organisations selling WiFi tracking technologies to provide a global opt-out list.⁶³⁹ FPF note on their website, however, that"

Owners of iOS 8 devices⁶⁴⁰ that wish to opt-out of Mobile Location Analytics can still do so by visiting the Smart Store Privacy Opt Out Page. However, since this opt out works by recognizing the MAC address of an opted-out device, in the case of iOS 8 devices, any such opt out will be reset when the device's MAC address changes.⁶⁴¹

This highlights another rights issue—that the PbD approach taken in the development of Apple devices, among others⁶⁴² prevents effective opting out without necessarily providing effective privacy. The ambient environment, much of which is rightfully untrusted, as anybody could silently set up a device capturing MAC addresses, leads hardware providers to make a value choice for data subjects. Whether opting out is possible given MAC randomisation is a research question in and of itself. Legally enforceable DNT signals may be required—something which raises many issues in and of themselves that I do not seek to unpack here, suffice to say that they would

(2016) 19(251) Privacy & Informatie.

⁶³⁵ Committee on Civil Liberties, Justice and Home Affairs (n 496).

⁶³⁶ College bescherming persoonsgegevens (n 611) 20. On opt-out registers in relation to the Internet of Things, see generally Lilian Edwards, 'Privacy, security and data protection in smart cities: A critical EU law perspective' (2016) 2 Eur. Data Prot. L. Rev. 28, 55.

⁶³⁷ Autoriteit Persoonsgegevens, 'Bluetrace beëindigt overtredingen wifi-tracking na optreden AP' [2017] Autoriteit Persoonsgegevens (Dutch Data Protection Authority) Website (<https://perma.cc/D78Q-WDBJ>).

⁶³⁸ Future of Privacy Forum, *Mobile Location Analytics Code of Conduct* (FPF 2013) (<https://perma.cc/LC4B-FHY5>)

⁶³⁹ <https://optout.smart-places.org/>.

⁶⁴⁰ iOS is the name of the operating system on Apple phones.

⁶⁴¹ Future of Privacy Forum, *About Mobile Location Analytics Technology* (FPF 2016).

⁶⁴² Meaning the MAC rotation practices documented in Martin and others (n 623).

require unprecedented coordination between the manufacturers of wireless tracking systems and those of mobile devices.⁶⁴³

3.1.4. Vignette of Vanished Rights 2: Apple's 'Siri' voice assistant

Voice assistants are now commonplace in a range of devices. Typically, these systems, including Microsoft's Cortana, Google's Assistant and Apple's Siri, work by recording and compressing audio data, processing it for transcription on the company's servers, and returning the transcript to the phone, where a local speech synthesis system may 'reply' to the user. The use of this approach has allowed unprecedented accuracy in speech recognition, as well as avoiding energy, resource and space-intensive processing on the terminal device. Many people use these technologies to activate device functionalities, or to dictate messages or documents.

Speech recognition has seen great improvements as a result of machine learning technology.⁶⁴⁴ Continuing challenges exist in areas such as noisy environments and multi-lingual input,⁶⁴⁵ but these are hoped to be soluble given the great amount of international voice data captured through the use of mobile devices and home assistants. As a consequence, such voice and estimated transcript data is very valuable to store for future training.

Some firms provide this recording data to data subjects upon request, while others do not. Google, for example, provide a tool where voice and audio data can be searched and managed.⁶⁴⁶ These can be seen as meeting their access obligations under European data protection law, although unlike many implementations of access rights, there does not appear to be a difference in these tools inside or outside the USA. Other firms, notably and as discussed here, Apple, despite providing a near-identical service to their competitors (in the sense that a voice assistant is provided on their devices), do not provide this data to data subjects automatically, nor do they provide such data upon explicit request under the Irish Data Protection Acts.⁶⁴⁷

Given the growing trend in collecting data for machine learning model training, much of it being sensitive, it seems worrying that large firms are managing to avoid to

⁶⁴³ The need for such coordination has been emphasised by the A29WP, see Article 29 Data Protection Working Party, *Opinion 13/2011 on Geolocation services on smart mobile devices* (WP 185) (n 620) 18.

⁶⁴⁴ See generally Li Deng and Xiao Li, 'Machine learning paradigms for speech recognition: An overview' (2013) 21(5) *IEEE Transactions on Audio, Speech, and Language Processing* 1060; Jayashree Padmanabhan and Melvin Jose Johnson Premkumar, 'Machine learning in automatic speech recognition: A survey' (2015) 32(4) *IETE Technical Review* 240.

⁶⁴⁵ Padmanabhan and Johnson Premkumar (n 644).

⁶⁴⁶ Google, 'Manage Google Voice & Audio Activity' (*Google Search Help*, 2017) (<https://perma.cc/BEJ3-PM3G>).

⁶⁴⁷ The author submitted a subject access request to Apple Distribution International, Ireland, which was denied. At the time, the GDPR was not in force, but the material reasons for denial and the legal grounds for making the request remain unchanged. The grounds for the denial are referred to in this section.

afford what appears to be a core responsibility under data protection law. This trend also seems to sit at tension with a common concern of individuals—that their devices are listening to them and targeting adverts on this basis. According to a 2018 survey by digital civil society organisation *doteveryone*, 7% of individuals in the UK *already* believe incredibly invasive data collection is taking place around them in relation to these types of functionalities: that conversations they have near their internet device are being collected, and 5% believe their eye movements when looking at the screen are collected.⁶⁴⁸ Despite this level of invasive collection for targeting advertising being rebuked by companies involved, and broadly put down to human biases ignoring irrelevant ads and only noticing those that are eerily relevant (by design or by chance) as well as the hugely targeted advertising ecosystem that *does* exist,⁶⁴⁹ access rights exist to allow users to check on such practices. As listening devices become more ubiquitous, this seems like an important facility for users to retain.

The argument made by Apple to the author concerning their non-provision of access rights is a highly useful launching point for a discussion of the tensions between data protection and the gathering of large datasets for personalisation and machine learning.

In correspondence (which can be found in Appendix 7.1) Apple cite PbD as the reason for this. Apple’s notion of PbD in relation to voice assistant data seems to hinge on three aspects.

Separation of identifiers Firstly, Apple claims that voice identifier data is divorced from the usual identifiers that Apple users are familiar with. While Google users log in with their account details, under which all their voice data are then listed, Apple generate device-specific identifiers that are separate from these identities.

When Siri is turned on, the device creates random identifiers for use with the voice recognition and Siri servers. These identifiers are used only within Siri and are utilized to improve the service. If Siri is subsequently turned off, the device will generate a new random identifier to be used if Siri is turned back on.⁶⁵⁰

Nevertheless, these are persistent identifiers. It appears that if the user never disables Siri in the device’s settings, as we might expect few users to do rather than simply

⁶⁴⁸ doteveryone, *People, Power and Technology: The 2018 Digital Attitudes Report* (doteveryone 2018) (<https://perma.cc/WT2C-SJ75>) 14.

⁶⁴⁹ See eg Gennie Gebhart and Jamie Williams, ‘Facebook Doesn’t Need To Listen Through Your Microphone To Serve You Creepy Ads’ (*Electronic Frontier Foundation (EFF) Deeplinks Blog*, 13th April 2018) (<https://perma.cc/QU7B-LA98>); Antonio García Martínez, ‘Facebook’s Not Listening Through Your Phone. It Doesn’t Have To’ (*Wired*, 11th October 2017) (<https://perma.cc/SLJ6-6YN9>).

⁶⁵⁰ Apple Inc, *iOS Security: iOS 10* (Apple Inc 2017) (<https://perma.cc/8EQE-TFW5>).

3. Data Protection's Lines, Blurred by Machine Learning

opting not to use it, the identifier persists throughout the lifetime of the device. Apple claimed in correspondence that they do not have a technical means to access the Siri identifier on the device, nor to search the data by identifier, as they have chosen not to build one.⁶⁵¹

Storage limitation Secondly, Apple claim that data usually have their linked identifiers scrubbed, and are eventually deleted after certain times have elapsed:

User voice recordings are saved for a six-month period so that the recognition system can utilize them to better understand the user's voice. After six months, another copy is saved, without its identifier, for use by Apple in improving and developing Siri for up to two years. A small sub-set of recordings, transcripts and associated data without identifiers may continue to be used by Apple for ongoing improvement and quality assurance of Siri beyond two years. Additionally, some recordings that reference music, sports teams and players, and businesses or points of interest are similarly saved for purposes of improving Siri.⁶⁵²

Data minimisation Thirdly, Apple claims that while Siri is able to recognise your name, it does this by sending such details from your phone each time Siri is used, until such a time where it has not been used for ten minutes, upon which it is deleted from the remote server.⁶⁵³

This conception of PbD comes with a number of issues.

Upon first glance, the above may seem like privacy-promoting design features. Yet there are significant conceptual flaws with each, as well as the entire system, that means while Apple currently find it difficult to access this data, re-identification would be possible, if not relatively trivial in some cases.

Firstly, refusing to build a database retrieval tool is no basis on which to refuse data subject rights. Retrieval is generally a standard feature of database systems. Indeed it is arguably their very purpose. In most cases, data controllers would have to proactively modify their systems in order to remove such functionality from standard database software.⁶⁵⁴

⁶⁵¹ '[W]e have not built any tool that allows us to retrieve this data'; email from Apple Distribution International to author (3 August 2017), see Appendix 7.1 of this thesis.

⁶⁵² Apple Inc, *iOS Security: iOS 10* (n 650) 50.

⁶⁵³ *ibid.*

⁶⁵⁴ Incidentally Apple argued in correspondence with the author that Siri data was not stored in a 'filing system', citing Article 2 of the GDPR on material scope. However exemption for data which do not 'form part of a filing system' is explicitly intended to apply only to data not processed by automated means. It would be unlikely that this line of argument would find much traction with regulators or in courts. At the time of writing the firm is not moving on its view.

Secondly, refusing to make the device identifier accessible to the data subject through the design of the software whilst still enabling it to be transmitted regularly to the data controller serves little practical purpose other than obstructing the data subject's ability to verify it is indeed them requesting the data.⁶⁵⁵ Indeed, this seems to be doing more to stand in the way of data protection rights than provide privacy by design. The GDPR is quite clear that such identifiers would be considered associated with a natural person, noting that 'online identifiers provided by their devices', including those provided by RFID,⁶⁵⁶ may either directly enable profiling or identification, or may do so indirectly, such as in combination 'with other information received by the servers'.⁶⁵⁷

Thirdly, while Apple notes that they do not permanently save the name you provide on the server, they do save many kinds of information of similar or even greater use in re-identification alongside your identifier. Indeed, Apple note that because it is onerous to send details such as relationships with family members, reminders, and playlists to the server each time a Siri session is started (and would likely introduce unwanted lag and/or data use), they send those initially, and store them there. Even if we were to accept that a device-specific identifier was not personal data (despite the rulings surrounding MAC addresses and even dynamic IP addresses), a list of their contacts and their relations to you is relatively trivial even for non experts to use to re-identify individuals by using easily accessible data sources, like social media. It seems similarly likely that simple re-identification attacks could be formulated against things such as reminders, particularly as they often mention the names of organisations or individuals.

Fourthly, a significant body of research has demonstrated that individuals can be re-identified and clustered by voiceprints alone, which have such re-identification potential that they are being used and proposed for biometric authentication.⁶⁵⁸ Apple themselves even possess several patents in this area from their own in-house research activities⁶⁵⁹ Even based on text transcripts without the voice data, researchers have

⁶⁵⁵ There can be useful reasons for obscuring data from both the user and the controller at a *hardware* level—*secure enclaves*, such as those that enable fingerprint scanning locally without making the verification data directly accessible to the rest of the system, work in this way.

⁶⁵⁶ Which being largely imperceptible, are similarly inaccessible to the average data subject. See further Edwards, 'Privacy, security and data protection in smart cities: A critical EU law perspective' (n 636).

⁶⁵⁷ GDPR, recital 30.

⁶⁵⁸ See eg N Dehak, P J Kenny, R Dehak, P Dumouchel and P Ouellet, 'Front-End Factor Analysis for Speaker Verification' (2011) 19(4) IEEE Trans. Audio Speech Lang. Processing 788. For opposing work on systems attempting to dodge re-identification, cf Federico Alegre, Giovanni Soldi, Nicholas Evans, Benoit Fauve and Jasmin Liu, 'Evasion and obfuscation in speaker recognition surveillance and forensics' in *2nd International Workshop on Biometrics and Forensics* (IEEE 2014) DOI: 10/cwnx.

⁶⁵⁹ See eg Jerome R Bellegarda and Kim E A Silverman, 'Fast, language-independent method for user authentication by voice' (*US Patent no 9218809*, 2015) (<https://patents.google.com/patent/US9218809B2/>); Adam J Cheyer, 'Device access using voice authentication' (*US Patent no 9262612*, 2016) (<https://patents>).

3. Data Protection's Lines, Blurred by Machine Learning

demonstrated attacks that can re-identify or cluster individuals stylometrically, based on the words and grammar they use.⁶⁶⁰

Compounding this, it is not just *how* things are said, but *what* is being said. Sensitive data can be said and held in textual form. How to redact terms that might disclose sensitive data is an active field of research.⁶⁶¹ This is very challenging even when the forms of text are relatively standardised, such as in medical documents⁶⁶²—standardisation not present in messages or other spoken interactions. As a recent review notes, '[g]eneral-purpose privacy solutions for plain text are scarce and they only focus on the protection of sensitive terms, which are assumed to be manually identified beforehand'.⁶⁶³ These systems have not been developed with conversation transcripts in mind; it is unclear that there are effective privacy mechanisms in place here that would defend against re-identification. Furthermore, sensitive data is likely to be recorded, including special categories of data under the GDPR, such as political opinions.⁶⁶⁴ Without guarantees that private and re-identifiable parts of a conversation have been redacted, which seem technically difficult, if not currently impossible, to provide, little assurance can be given.

3.1.5. Rescuing Data Protection by Design

If DPbD risks taking away rights, as in the cases illustrated above, how might we rectify or ameliorate this? How might we put the *data protection* back into *data protection by design*? Here, I propose some approaches that might help do this, assess their possibilities and pitfalls, and place them in legal context.

3.1.5.1. Parallel Systems for Data Rights

One set of options would be to maintain parallel systems with the explicit purpose of upholding these rights. Here, I outline two main types of these systems in legal and technical context: systems designed to retain data to provide access and better enable

google.com/patent/US9262612B2/); Allen P Haughay, 'User profiling for voice input processing' (*US Patent no 9633660*, 2017) (<https://patents.google.com/patent/US9190062B2/>).

⁶⁶⁰ S Afroz, A C Islam, A Stoleran, R Greenstadt and D McCoy, 'Doppelgänger Finder: Taking Stylometry to the Underground' in *2014 IEEE Symposium on Security and Privacy* (IEEE 2014) DOI: 10/cwnz.

⁶⁶¹ David Sánchez and Montserrat Batet, 'Toward sensitive document release with privacy guarantees' (2017) 59 *Engineering Applications of Artificial Intelligence* 23.

⁶⁶² See eg Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen and Matthew H Samore, 'Automatic de-identification of textual documents in the electronic health record: a review of recent research' (2010) 10(1) *BMC Medical Research Methodology* 70.

⁶⁶³ Sánchez and Batet, 'Toward sensitive document release with privacy guarantees' (n 661) 24.

⁶⁶⁴ GDPR, art 9.

erasure and objection, and systems designed to process additional data, which may be provided by the data subject, to make re-identification possible.

3.1.5.1.1. Obligations to Retain Data In *Rijkeboer*⁶⁶⁵ the CJEU was referred a question relating to a case where Mr Rijkeboer, a Dutch citizen, asked the Mayor and Executive Board of Rotterdam to provide him with details of the third parties to which any information relating to him held by the municipality had been communicated to.⁶⁶⁶ In Mr Rijkeboer's case, the data controller had replied positively but partially, providing only information relating to the previous year, as Dutch law and practice provided that the data from the year preceding had been wiped. The question to the court was whether, in the absence of a timeframe provided within the access rights of the DPD, Member States could impose deletion of such data—which was in a sense *metadata* about the data held on Mr Rijkeboer⁶⁶⁷—after a certain period of time—meaning that such access rights could not refer to data outside this time period.

The Mayor and Executive Board of Rotterdam, the United Kingdom, the Czech, Spanish and Dutch governments submitted that the right of access 'exists only in the present and not in the past', while the Greek government and the European Commission submitted that it applies not only to the present but also extends into the past.⁶⁶⁸ The court ruled that such a right must necessarily relate to the past to ensure the practical effect of access, erasure and rectification provisions,⁶⁶⁹ that the exact time limitation was up to further Member State rule-making, but that a period of one year alone does 'not constitute a fair balance of the interest and obligation at issue' unless it can be shown that anything longer would lead to an 'excessive burden' on the controller.⁶⁷⁰ Indeed, any time limit upon this metadata should constitute 'a fair balance between, on the one hand, the interest of the data subject in protecting his privacy, in particular by way of his rights to object and to bring legal proceedings and, on the other, the burden which the obligation to store that information represents for the controller.'⁶⁷¹

This ruling is pertinent to the current discussion, as it directly places the question of data subjects' rights against what the court described as the 'burden'⁶⁷² that data storage places on the controller—a burden which consists increasingly of securing this data against adversaries, rather than just the simple cost of storage media. The CJEU

⁶⁶⁵ *Rijkeboer* (n 413).

⁶⁶⁶ Under DPD, ¶ 12.

⁶⁶⁷ Note that the Court in *Rijkeboer* (n 413) does not refer to this as metadata (but as another category of data in contrast to so-called 'basic data'), I do so here for explanatory purposes.

⁶⁶⁸ *ibid* ¶¶ 37–39.

⁶⁶⁹ *ibid* ¶ 54.

⁶⁷⁰ *ibid* ¶ 66.

⁶⁷¹ *ibid* ¶ 64.

⁶⁷² *ibid* ¶ 64.

3. Data Protection's Lines, Blurred by Machine Learning

acknowledged that this was a trade-off that the DPD did not contemplate explicitly: the same can be said of the GDPR.

Another relevant point from this ruling is the distinction made by Court between two types of data in light of the right of access.⁶⁷³ Firstly, that the 'basic data', used for the functionality of local service provision, was being stored for a longer period than the data regarding the transfers (which of course may be sensitive to the data subject), was noted to be a source of the unfair balance that had been struck by the Rotterdam Mayor and Executive Board.⁶⁷⁴ Put differently, one could say the controller adopted a different retention policy for 'content data' (ie the actual personal data such as individuals' names) as opposed to 'metadata' (eg information relating to how the personal data was used and its source). This has an interesting, although not exact, parallel to some alleged PETs. In these technologies, it is also possible to distinguish between different types of data; the full, potentially identifiable data collected, and the transformed data which is now more difficult to link to data subjects. The former is erased after a certain timeframe,⁶⁷⁵ often at the time that it is transformed into the latter for retention. Is this erasure a 'fair balance'?

Distinguishing different types of information for the purposes of data management is a common industry practice. Yet the ways in which data are classified within organisations do not always have neat analogues the legal framework. Take the Siri identifiers discussed above.⁶⁷⁶ Siri identifiers clearly single out a data subject, as they are persistent identifiers that link to a device typically only used by one person. One of the major purposes of this system is to deliver a personalised voice assistant to a data subject. As a result, Article 11(1) GDPR, which relieves data controllers from having to 'maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying' with the Regulation (notably accommodating data subject rights), does not apply. Data suitable for use in some context to identify the data subject is *already being processed* in the form of the Siri identifier; it is simply that Apple refuses to repurpose that data to allow for the provision of access rights. And even if it *were* to apply, Article 11(2) still enables data subjects to have their data subject rights accommodated upon providing additional information that does allow the controller to (re-) identify them.

However, the Regulation does not seem to contemplate that technological develop-

⁶⁷³ *Rijkeboer* (n 413) 42 et seq.

⁶⁷⁴ *ibid* 42 et seq.

⁶⁷⁵ In the case of Siri, Apple *further* de-identifies this data after a six month period, noting that 'User voice recordings are saved for a six-month period so that the recognition system can utilize them to better understand the user's voice. After six months, another copy is saved, without its identifier, for use by Apple in improving and developing Siri for up to two years.'. See Apple Inc, *iOS Security: iOS 10* (n 650) 50.

⁶⁷⁶ Section 3.1.4.

ments have allowed identification of a data subject for the purpose of service delivery and data processing, but not for the purposes of data access. This is worrying, particularly given the increase in development of asymmetric technologies with these characteristics, such as zero-knowledge proofs, by which one party (a prover) can prove to another (the verifier) that they know some information x , without conveying any information apart from the fact that they know x .⁶⁷⁷ Such technologies are likely to throw further strain on this assumption in European data protection law, particularly as machine learning-based personalisation technologies might become delivered with the aid of cryptographic authentication.

If there is a split form of identification and singling-out as described above, how might a data controller use a parallel system to augment the identification process being undertaken for service delivery to also allow data access? Apple IDs could be stored alongside Siri identifiers in a separate database. While Recital 64 of the GDPR does note that a controller ‘should not retain personal data for the sole purpose of being able to react to potential requests’, here the controller already holds both sets of personal data, and only needs to establish a link between them. Asking controllers to purposively make it difficult to consistently find a data subject across many datasets held seems problematic in light of the practical challenges of GDPR implementation. A mechanism could also be implemented on the device to obtain the identifiers used.

The core question that relates to these approaches is of security, and in particular, I suspect, national security. A centralised list in the first case presents a significantly heightened re-identification risk were *attackers* to gain access to this data. Yet both options jeopardise what one might suspect to be among Apple’s deeper aims—to claim their hands are tied when faced with law enforcement or intelligence services requests, as they have done publicly before.⁶⁷⁸

While these approaches rely on burying data in a haystack, it may also be possible for Apple to provide this data to users in a form only *they* can access, using encryption techniques. This might be done, for example, by devices transmitting data not only to Apple, but *also* (likely optionally) to a server or device the user controls, including controlling the encryption key of.⁶⁷⁹ In this case, the data controller would

⁶⁷⁷ Zero knowledge proofs of identity have been theorised for some decades already: see Uriel Feige, Amos Fiat and Adi Shamir, ‘Zero-knowledge proofs of identity’ (1988) 1(2) *Journal of Cryptology* 77. For a classic and somewhat seminal layman’s description of the intuition behind the technology, see Jean-Jacques Quisquater, Louis Guillou, Marie Annick and Tom Berson, ‘How to Explain Zero-knowledge Protocols to Your Children’ in *Proceedings on Advances in Cryptology (CRYPTO ’89)* (Springer 1989). Interest is growing of late due to more powerful and connected devices, which is evidenced in a number of official international standardisation efforts around zero knowledge protocols.

⁶⁷⁸ Karl Stephan, ‘Apple Versus the Feds: How a smartphone stymied the FBI’ (2017) 6(2) *IEEE Consumer Electronics Mag.* 103 DOI: 10/cwvf.

⁶⁷⁹ Such a device might, for example, be a personal data container. See eg Andy Crabtree, Tom Lodge, James Colley, Chris Greenhalgh, Richard Mortier and Hamed Haddadi, ‘Enabling the new economic actor: data

not be retaining the data in a form they could access to provide upon request, but instead providing portability from the outset. Indeed, users might find it useful to have a repository of speech data and transcripts in order to quickly train any new system, were they to change providers. They might also wish to hold it for autobiographical or 'self-logging' reasons.⁶⁸⁰ If DPbD means access and portability 'by design', these are feasible design solutions that could form part of the strategy from the outset. This may also allow a user the effective right of erasure; (hashes of) the data they hold could be automatically compared with the de-identified database, and the matches removed.

Just as *Rijkeboer* forced the data controller to rethink their data retention process, it seems feasible that future rulings—just as the drafting of the GDPR has⁶⁸¹—could also, inventively, take aim *further upstream*, at the 'fair balances' being struck in the design process.

3.1.5.1.2. Acquiring Additional Information Re-identifying data with an acceptable degree of certainty in order to exercise data protection rights may be difficult in practice for any data controller practicing certain types of DPbD, regardless of intention. The GDPR recognises this in Articles 11 and 12(2), which exempt the data controller from having to accommodate data subject rights if it can demonstrate it is not in a position to identify the data subject. Article 11(2) however, grants data subjects the ability to provide additional data to enable such (re-)identification, something not every data subject might be inclined or able to do.⁶⁸² The final call though, seems to be in the hands of the controller. Pursuant to Article 12(2) *in fine*, the data controller still has an opportunity to demonstrate not being in a position to (re-)identify, even after being provided with additional information by the data subject.

Having said all that, it would still require a considerable burden of proof to adequately demonstrate reidentification is not possible, even despite additional information being provided by the data subject. This does not only appear from the GDPR's general emphasis on accountability⁶⁸³ and weightier focus on data controllers' responsibilities, but also manifests itself through Recital 57. This Recital notes that while the data controller 'should not be obliged to acquire additional information in order to identify the data subject' to comply with the regulation, they 'should not refuse to take additional information provided by the data subject in order to support the exercise

protection, the digital economy, and the Databox' (2016) 20(6) Personal and Ubiquitous Computing 947 DOI: 10/f9b3hp.

⁶⁸⁰ See generally Gina Neff and Dawn Nafus, *Self-Tracking* (MIT Press 2016).

⁶⁸¹ See also section 2.4.

⁶⁸² Indeed, as data is increasingly processed by cloud compute services, individuals may not store or retain the copies themselves needed to identify them, particularly when this data is not used directly by data subjects.

⁶⁸³ GDPR, art 5(2).

or his or her rights'.⁶⁸⁴ Taking a step back, it is of course important to emphasise that an (alleged) inability to fully accommodate data subject rights cannot be exploited to evade data protection law altogether, and that all other provisions (notably those in Article 5 and 6) still apply in full.⁶⁸⁵

An unanswered question remains—such acquired or volunteered additional information still requires a re-identification process that while very possible, may not be straightforward to the data controller to undertake. Indeed, data controllers may not have expertise in this space, particularly when it does not form part of their core processing activities. While Recital 57 shines little light on this, Recital 26 provides some guidance, suggesting that factors such as cost, time, available and emerging technology should be taken into consideration.⁶⁸⁶

This provides an interesting avenue for a policy intervention—a possibly controversial one—that could support data subjects' rights. While theoretical attacks for re-identification are often possible, and would likely undermine the privacy-by-design approaches taken above, there is a valid argument about whether these technologies are 'available' (in the context of Recital 26). While they might be available to attackers, and traded, like stolen data, on shady online markets, this creates an imbalance between the deployable technologies available to data controllers and those available to their adversaries. Is there an obligation on data controllers to develop (or to procure from security companies) state-of-the-art re-identification tools in order to make data subject rights possible?

There is a parallel here with other examples, albeit not all in the EU, in which certain individuals are owed redress by an organisation who lacks the means of identifying them. After a U.S. financial lender, Ally Financial, was found to have racially discriminatory car loan pricing, they were ordered to use census data to estimate which of their borrowers were Black, Hispanic or Asian in order to (imperfectly) identify the rightful recipients of compensation.⁶⁸⁷ Similar efforts might be beneficial in the case of data breaches, where publicly available information could be mined in order to identify a means of contacting affected individuals who are otherwise only known to the controller by their browsing history, device fingerprint or other data.

Data protection law, in an attempt to be technologically neutral, is silent on impos-

⁶⁸⁴ GDPR, recital 57.

⁶⁸⁵ Article 29 Data Protection Working Party, *Opinion 1/2010 on the Concepts of 'Controller' and 'Processor.'* (WP 169) (2010).

⁶⁸⁶ See GDPR, recital 26, which states that '[t]o ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.'

⁶⁸⁷ Annamaria Andriotis and Rachel Louise Ensign, 'U.S. Uses Race Test to Decide Who to Pay in Ally Auto-Loan Pact' (*Wall Street Journal*, 29th October 2015) (<https://perma.cc/NBH8-ERDS>).

3. Data Protection's Lines, Blurred by Machine Learning

ing specific innovation requirements on data controllers—which is probably a good thing, as mandating technological advancement through legislating it seems like a misguided idea.⁶⁸⁸ But were governments, academia or civil society to develop and make re-identification tools for high dimensional data publicly available, with a codebase compatible with many types of commercial systems, it would be hard to deny these technologies were 'available' in the sense of Recital 26. Additionally, the possibility for certification bodies outlined in the GDPR may provide a further avenue for keeping up-to-speed on the state of the art technologies in this space.⁶⁸⁹ Yet this comes with its own security risks. Not only are these tools then available to attackers, but they may even be installed and calibrated on the systems that data is being illicitly obtained from, leaving adversaries a little like the proverbial 'kid in a candy store'.

When these re-identification mechanisms are already designed however, and out in the published research literature, 'putting a lid on them' would appear to be a poor policy approach. Even where the codebase is scrappy and unreliable, these are precisely the types of tools that 'script kiddie' adversaries are used to working with. Imagining that more usable tools only serves to help attackers probably itself underestimates adversaries' existing capacity to use and generate knowledge to valorise stolen personal data, as well as understates these tools' benefits in giving data subjects more control over the data they are entitled to legal rights over. Indeed, making them more usable may not vastly increase the capabilities of attackers that were always willing to string together unreliable code, and may primarily serve to empower data subjects to manage risks relating to them.

3.1.5.2. Making Trade-offs Accountable

Given the value-laden nature of these trade-offs, it is important that they are made in an explicit way, with care and with rigour. As it stands, the GDPR, being extremely vague about what DPbD means, does not acknowledge either in recitals or the enacting terms the existence of trade-offs within design approaches. Given these trade-offs, as I have shown, involve fundamental rights of data subjects, this seems unacceptable.⁶⁹⁰

Data protection impact assessments (DPIAs)⁶⁹¹ are positioned as a potentially apt point in the compliance process to consider the trade-offs present when employing

⁶⁸⁸ This is not to say that the state should not have a role in steering innovation, or strategically funding particular areas—indeed, it often has—but that innovation policy is more complex than imposing a statutory requirement. See, as a lay translation of the innovation studies literature, Mariana Mazzucato, *The entrepreneurial state: Debunking public vs. private sector myths* (Anthem Press 2015).

⁶⁸⁹ GDPR, art 42.

⁶⁹⁰ The very ability for an individual to have access to their personal data forms an explicit part of the fundamental right to the protection of personal data: see Charter, art 8(2).

⁶⁹¹ GDPR, art 35.

DPbD strategies. DPIAs are the main form of preemptive analysis and documentation requirement in the GDPR, taking particular aim at high risk processing.⁶⁹² The GDPR explicitly, albeit in the recitals, notes that the ‘risk to the rights and freedoms of natural persons, of varying likelihood and severity, may result from personal data processing which could lead to physical, material or non-material damage, in particular: [...] where data subjects might be deprived of their rights and freedoms or prevented from exercising control over their personal data.’⁶⁹³ Accordingly, the A29WP identifies, as one of the criteria leading to high risks to data subjects, situations where ‘the processing in itself “prevents data subjects from exercising a right or using a service or a contract”’.⁶⁹⁴

Yet in the same guidance, it is worth noting that DPbD gets only a fleeting mention as another pre-emptive approach comparable to DPIA; the A29WP remains silent on including DPbD itself as a topic within DPIAs.⁶⁹⁵ It is furthermore easy to see how DPbD measures such as those discussed could be seen as one of several ‘measures envisaged to address the risks, including safeguards, security measures and mechanisms’.⁶⁹⁶ As processing undertaken in response to otherwise risky processing with the intention of decreasing that risk, they might escape the scrutiny applied to the original concern. While an infinitely recursive DPIA is highly undesirable, so is one that lacks appropriate reflexivity.

While a DPIA might, potentially and with further clarification, provide a venue for considering trade-offs, this approach has a number of limitations owing in particular to the weakening of certain key provisions in the final text of the GDPR.⁶⁹⁷ While the requirement to ‘seek the views of data subjects or their representatives’ during the DPIA process suggests that those individuals affected can articulate their views about appropriate trade-offs, this obligation is limited. It is only required ‘where appropriate’ and ‘without prejudice to the protection of commercial or public interests’. The exemption from consulting data subjects where it might affect ‘the security of the processing operations’, presents yet another situation in which protection of data through obscurity could excuse the pursuit of other substantive data protection obligations. As a result of these limitations, such consultations may often in practice constitute a form-filling task, particularly as these views to be sought are not grounded in any particular task or question, and are not (as I discuss further below) required to be published or publicised.

⁶⁹² See generally Binns, ‘Data protection impact assessments: A meta-regulatory approach’ (n 332).

⁶⁹³ GDPR, recital 75.

⁶⁹⁴ Article 29 Data Protection Working Party, *Data Protection Impact Assessment Guidelines* (n 534) 11.

⁶⁹⁵ *ibid* 14.

⁶⁹⁶ GDPR, art 35(7)(d).

⁶⁹⁷ See generally Binns, ‘Data protection impact assessments: A meta-regulatory approach’ (n 332).

3.1.5.3. Information Rights around Privacy Architectures

As described in the vignettes above,⁶⁹⁸ we can increasingly locate examples where data subjects' personal data is being processed without the accompanying data subject rights effectively being enabled. Yet it seems rare for data subjects to be informed before the time of collection or processing that such rights will not apply. Where they are, claims seem highly generalised. Apple's Privacy Policy, for example, simply states that it 'may decline to process [access] requests that are frivolous/vexatious, jeopardize the privacy of others, are extremely impractical, or for which access is not otherwise required by local law.'⁶⁹⁹ Which data will be 'extremely impractical' to exercise rights over? Which will be considered to 'jeopardize the privacy of others'? Without this information, it seems unclear that a proper evaluation could—or should—be made by a data subject as to whether she wishes to entrust her personal data to such a controller.

Must a data controller, explicitly and without request at the time data are obtained, warn a data subject that the rights they might expect do *not* exist? This would seem critical if, as data protection law expects, data subjects are to play a part in managing the risks in accordance with their own preferences. There is a requirement to provide 'information necessary to ensure fair and transparent processing', including 'the existence of the right to request from the controller access to and rectification or erasure of personal data'.⁷⁰⁰ Yet it is unclear whether this is a provision that requires the existence of these rights in a general sense—an awareness-raising measure, as well as one seeking to provide logistical support (eg through pointing to the relevant controller contact details)—or whether this is an existence of these rights in applied context, considering each type of data processed by the controller. Considering that Article 11(2) contemplates times when there might be no 'existence' of these rights, it makes sense that this requirement would not apply in those cases (as to do so would be to mislead). Does, however, this mean that data controllers would have to *invert* the obligation, and explicitly tell data subjects that their rights will *not* be honoured? This is unfortunately less clear. The A29WP do not, unfortunately, address this in their guidelines on transparency.⁷⁰¹ Yet, in light of the overarching transparency and fairness principles in Article 5(1), it seems like such a reading is justified.

Other parts of the law are more clear, however. It is apparent that there *is* an obligation on data controllers to provide the reasons for their non-fulfilment of a specific

⁶⁹⁸ See section 3.1.3 and section 3.1.4.

⁶⁹⁹ Apple Inc, 'Privacy Policy' (apple.com, 19th November 2017) (<http://perma.cc/3DC2-M7Z5>).

⁷⁰⁰ GDPR, art 13(2)(b); GDPR, art 14(2)(c).

⁷⁰¹ Article 29 Data Protection Working Party, *Guidelines on Transparency under Regulation 2016/679 (wp260rev.01)* (2018).

access request after it has been made. In contrast to the unclear scope of *ex ante* information requirements,⁷⁰² relevant *ex post* information requirements expect controllers to provide more detailed information on when subject rights are not available and why.⁷⁰³

Despite this, there is no explicit hook in Articles 13–14 for a data controller to provide information *ex ante* about the DPbD measures that might be restricting such rights so that a data subject might assess such provisions and their appropriateness, nor provide information on the safeguards being applied to their data that might affect their ability to exercise their rights. Given the importance of these rights to the data protection regime as a whole, this seems problematic in relation to the transparency and accountability principles.⁷⁰⁴

I argue however that such a requirement *can* be read into the *ex ante* obligation to provide ‘meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject’ of automated decision making.⁷⁰⁵ Commentators have historically viewed the potential of automated decision rights in relation to ‘algorithmic accountability’ discussions, through the lens of ‘decisions’ individuals encounter in their day-to-day-lives such as credit scoring or behavioural targeting.⁷⁰⁶ Yet in relation to the envisaged removal of fundamental rights using automated processing, I argue that an automated decision (‘which may include a measure’⁷⁰⁷) could also be considered in relation to processing that happens *internally*, within a data controller or processor.

These rights, as already discussed in this thesis, have been considered strongly restricted by both a restriction to be ‘solely’ automated, and to trigger ‘legal’ or ‘similarly significant’ effects on individuals.⁷⁰⁸ Yet there is strong reason to believe that the systems being discussed here meet both conditions, and therefore do not fall foul of the

⁷⁰² GDPR, art 13; GDPR, art 14.

⁷⁰³ See GDPR, recital 59, stating that ‘[t]he controller should be obliged to respond to requests from the data subject without undue delay and at the latest within one month and to give reasons where the controller does not intend to comply with any such requests.’; see further GDPR, art 12(4), stating if ‘the controller does not take action on the request of the data subject, the controller shall inform the data subject without delay and at the latest within one month of receipt of the request of the reasons for not taking action’.

⁷⁰⁴ It could be argued that fundamental rights are at stake here in some situations, not just the right to data protection (Charter, art 8) but also the right to non-discrimination (Charter, art 21) and to freedom of expression (Charter, art 11).

⁷⁰⁵ GDPR, art 13(2)(f); GDPR, art 14(2)(g).

⁷⁰⁶ See eg Bygrave, ‘Automated Profiling: Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling’ (n 345); Mireille Hildebrandt, ‘The Dawn of a Critical Transparency Right for the Profiling Era’ in Jacques Bus, Malcolm Crompton, Mireille Hildebrandt and George Metakides (eds), *Digital Enlightenment Yearbook* (IOS Press 2012) DOI: 10/cwpm; Edwards and Veale, ‘Slave to the Algorithm?’ (n 79). See also the algorithmic war-stories in section 1.4.

⁷⁰⁷ GDPR, recital 71.

⁷⁰⁸ See section 2.2.1.1.

3. Data Protection's Lines, Blurred by Machine Learning

narrow applicability of these rights.

Firstly, privacy enhancing technologies rarely have humans in-the-loop after their initial setup—usually this would undermine mechanisms reducing information disclosure—and as a result, we can broadly think of these technologies as ‘solely’ automated. Secondly, the removal of rights would arguably have both a ‘legal effect’, in the sense of changing a data subject’s position with respect to Article 11, and a ‘similarly [significant]’ effect, impacting on fundamental rights and freedoms.

As there appear to be grounds to meet this condition, it must lastly be considered whether or not a discernable ‘decision’ has been made. The most clear indication that it has in the case of DPbD is that Recital 71 specifically includes that the scope of automated decisions ‘may include a measure’—the precise terminology in Article 25(1) describing DPbD as ‘technical and organisational *measures* necessary to ensure, for the processing concerned, that this Regulation is implemented’ [emphasis added]. Some might say that these measures happened at the moment of system design, not at the point of processing, and therefore, not being solely automated nor affecting a single data subject at that point, no information obligation exists. Yet to apply this reasoning to profiling systems, such as behavioural advertising, would be absurd. While at a mechanical level, visiting a webpage might trigger the application of a pre-built profile to deliver advertising,⁷⁰⁹ the ‘logic involved’ would presumably not (and seemingly not in the eyes of the A29WP⁷¹⁰) be restricted to the last leg alone—that a user requested online components, which matched a browser fingerprint to a profile accessed a database, and therefore was provided specific content—but would refer to the *broader system* insofar as it was relevant to the final decision. This would include the construction of the profile in question. In a similar manner, rights to understand DPbD systems which are applied automatically would presumably have some broader, systemic notion applicable to them as well.

In the case of a ‘measure’ such as DPbD, it might be hard to imagine what a right to a human ‘in-the-loop’, the core remedy offered in Article 22 that Articles 13–15 refer to, would look like in this situation. Yet these information rights are not explicitly fully restricted by the compatibility of the remedy in a separate article. Indeed, I note that not only does the terminology in Articles 13–15 refer to ‘automated decision-making’, without either of the conditions in Article 22, it also counsels that it is ‘at least’, not only, in the context of Article 22 that these rights trigger, opening the door for less restrictive judicial interpretations in the future.

⁷⁰⁹ The A29WP indeed contemplate the possibility of advertising meeting the triggering conditions for GDPR, art 22. See Article 29 Data Protection Working Party, *ADM Guidelines* (n 2); for analysis, see Veale and Edwards, ‘Clarity, Surprises, and Further Questions’ (n 254).

⁷¹⁰ Article 29 Data Protection Working Party, *ADM Guidelines* (n 2).

The effect of this reading of automated decision information rights on DPbD measures which prevent the effective exercise of other rights would be twofold. It would firstly oblige controllers to provide ‘meaningful information about the [...] significance and envisaged consequences’ of such processing—the loss of data protection rights. They would have to do this *ex ante*—the A29WP has recently taken the view that the ‘meaningful information’ rights in Articles 13–14 should provide identical information to those in Article 15.⁷¹¹ At *minimum*, this provision would have the same effect I argue is present in Article 13(2)(b) and 14(2)(c) above, reinforcing our reading of the GDPR that to inform data subjects of these lack of rights is mandatory. A more generous reading could even see it go beyond this. The ‘consequences’ of the loss of data protection rights include a loss of control, and as such this might entail a discussion of the re-identification risk were such data to be accessed without authorisation. Insofar as ‘envisaged’ is understood as ‘intended’ rather than ‘foreseen’,⁷¹² it could be countered that the data controller does not ‘intend’ a data breach, and therefore would not be required to inform data subjects about its potential consequences. Yet given that such a breach could be highly damaging to data subjects, it is likely to trigger the separate ‘significance’ requirement, even were it to dodge the ‘envisaged’ one.

The second consequence relates to the ‘meaningful information about the logic involved’ requirement. This gets us closer to an obligation on the data controller to provide information about the extent, form and structure of relevant safeguards in a way that can be assessed by the data subject—or indeed, given these are *ex ante* information rights not requiring an existing data subject to trigger, by interested parties more broadly. The Oxford English Dictionary defines ‘logic’ as ‘a system or set of principles underlying the arrangements of elements in a computer or electronic device so as to perform a specified task.’ Where the task is partial de-identification or some other computational transformation to render such data difficult to single out, and thus deprive and individual data subject of certain data protection rights, this would indicate that a basic—and importantly, a ‘meaningful’—schematic would be provided. The ‘meaningful’ condition, one of the few changes to these rights from the DPD,⁷¹³ obliges information about this logic to relate to the data subject in a useful way—but given a lack of detailed requirements written in the GDPR, we will all likely have to wait for this right to be tested, and to potentially be referred to the CJEU, to understand how far it will take data subjects.

Despite a lack of detailed requirements relating to information rights, the account-

⁷¹¹ See Veale and Edwards, ‘Clarity, Surprises, and Further Questions’ (n 254).

⁷¹² The German version of the law is perhaps better translated in this way. See Wachter, Mittelstadt and Floridi (n 201) 84.

⁷¹³ cf GDPR, arts 12–15, with DPD, ¶ 12.

3. Data Protection's Lines, Blurred by Machine Learning

ability provisions of the GDPR⁷¹⁴ mean that, data controllers *are* expected to be able to adequately demonstrate compliance with *all* GDPR provisions, which includes security and DPbD obligations. This is not a requirement for such demonstrations to be user-facing, but that such a demonstration must be possible to eg the regulator, when or if they are asked.

As discussed above,⁷¹⁵ DPIAs might be an important venue for demonstrating this compliance and hammering out the trade-offs faced. As it stands however, they do not seem to be a reliable transparency mechanism—there is no obligation to publish these documents under the GDPR, and indeed industry opinion is highly opposed to such an obligation, usually on grounds of their potential to contain trade secrets and proprietary information.⁷¹⁶ When passed to a DPA as part of prior consultation,⁷¹⁷ the documents may become subject to local FOI laws, but given that data controllers can avoid prior consultation by claiming they have mitigated the risk, it is yet to be seen how common prior consultation will be in practice.

The problem with the lack of publishing of this information does not relate to an imaginary world where engaged data subjects pore over the minutiae of DPIAs, but in general the lack of rigorous scrutiny expected of a pluralist society afforded to organisational or technical approaches to privacy. Oversight is unlikely to be useful if only provided at an individual level. As discussed above,⁷¹⁸ just as individuals suffer from consent fatigue, many of the solutions for the increasingly complex processing ecosystem today run the risk of a 'transparency fallacy', where the responsibility for obtaining and digesting complex information about computational systems falls, un-

⁷¹⁴ GDPR, art 5(2); GDPR, art 24(1).

⁷¹⁵ See section 3.1.5.2.

⁷¹⁶ See the responses to the draft version of Article 29 Data Protection Working Party, *ADM Guidelines* (n 2) in the response to an FOI request from the author to the European Commission, DG Justice and Consumers (https://www.asktheeu.org/en/request/a29wp_data_protection_impact_ass), in particular the enclosed response from the trade body for most of the large platforms, *DIGITALEUROPE* expressing that view, among others. Indeed, trade secrets or intellectual property have been a traditional carve-out in the area of rights to 'logic of the processing'. Yet, according to Malgieri, 'if a conflict should arise between privacy rights of individuals and trade secret rights of businesses, privacy rights should prevail on trade secret rights.'. See Gianclaudio Malgieri, 'Trade Secrets v Personal Data: a possible solution for balancing rights' (2016) 6(2) *International Data Privacy Law* 102 DOI: 10.1093/idpl/ipv030, 103. GDPR, recital 63 does acknowledge that access rights should not adversely affect 'trade secrets or intellectual property and in particular the copyright protecting the software,' but also mentions that such arguments cannot be (ab)used to refuse access altogether. Similarly, the EU Trade Secrets Directive also notes that its provisions "should not affect the rights and obligations laid down" in the DPD. See 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, recital 35.

⁷¹⁷ GDPR, art 36(3)(g). A prior consultation must be carried out on the occasion that the data controller cannot mitigate the risks of a processing activity. Whether or not data controllers will self-declare a (subjective) inability to mitigate risks of their business models and allow the DPA to be in a position to veto it is highly unclear. This author feels that it will be a rare event indeed.

⁷¹⁸ See section 2.3.1.

helpfully, on the data subject. Instead, having third parties placed as beneficiaries of some information rights would be a useful future step. While DPAs have significantly increased powers to investigate data controllers, this usually happens only after a complaint has been raised.⁷¹⁹

It is difficult to raise a complaint about improper or ineffective applications of privacy or data protection by design without some insight into the system infrastructure—that is, until such systems fail in a large and noticeable way, at which point transparency is hardly a helpful remedy. In theory, high risk processing where risks cannot be sufficiently mitigated must involve consultation and prior authorisation with the responsible DPA⁷²⁰. In the vignettes above, surrounding Apple and TfL, assessment here of trade-offs made during mitigation of risk was only really possible due to the voluntary publishing by Apple of an iOS Security White Paper⁷²¹ and the fact that as a public body, TfL can be forced to release documents through FOI law.⁷²²

Yet as the precise risks that have been discussed here occur *during this mitigation process*, the trade-offs are commonly ignored—because they present *separate* risks from the initial assessment of whether processing is ‘high-risk’ or not. They will often not be flagged and identified similarly. While some collective or representative use of rights is supported by the GDPR,⁷²³ these rights do *not* include any concerning information provision (instead, they concern remedies including complaint, judicial action and compensation). Aligning information rights with those that can understand, investigate and report potential breaches, or simply to publicly highlight the existence of state-of-the-art technologies that can better make the trade-offs between different aspects of privacy and data protection in the context of DPbD, seems like a requirement for the future, particularly as those datasets and surrounding data systems, especially those powering machine learning models, become more pervasive, invisible and complex.

⁷¹⁹ Indeed, given the resource limitations of DPAs, it is hard to see proactive investigations affecting anything but the most high profile actors. The UK’s Information Commissioner has noted that her office has a history of taking forward complaints even where there is no data subject mandating them. She made these comments in relation to national debate around whether the UK makes a derogation to incorporate GDPR, art 80(2) (on ‘supercomplaint’-like mechanisms. Footnote 485 elaborates on the super-complaint features of UK law. But even this remains very different from a solo, proactive investigation—and the way DPAs handle these is yet to be seen. See Information Commissioner’s Office, ‘The Information Commissioner’s Office (ICO) response to DCMS General Data Protection Regulation (GDPR) derogations call for views.’ (ico.org.uk, 10th May 2017) (<https://perma.cc/33X9-HPHE>) para 113.

⁷²⁰ GDPR, art 36.

⁷²¹ Apple Inc, *iOS Security: iOS 10* (n 650).

⁷²² Both Transport for London, *TfL WiFi Analytics Briefing Pack* (n 615) and Transport for London, *Insights from Wi-Fi Data: Proposed Pilot* (n 610) were a result of FOI disclosures.

⁷²³ GDPR, art 80(1), which envisage some role for bodies to exercise rights either mandated by a data subject or (optionally, subject to member state derogations) without them respectively. See further section 2.3.2.

3.1.6. Interim remarks

Splitting individuals from their rights might not be a deliberate undertaking by nefarious data controllers, but its consequences matter. This section has effectively shown that data controllers, particularly those building machine learning systems which can extract value from these non-identified datasets, can interpret provisions such as PbD and data minimisation in ways that seem intuitively harmful due to the range of trade-offs often made implicitly. While PbD was initially defined as a holistic concept, its well-rounded nature has been somewhat lost, and it is increasingly seen as a synonym for the formal PETs literature that takes reducing unwanted information disclosure as a primary if not sole goal: similarly issues of fairness, focussing on a mathematically tractable, single optimisation target. The approach to address this challenge, I believe, surrounds the processes used to make, justify and contest these decisions. The mechanisms described that might 'rescue' data protection by design within the existing framework all take aim at process, and in doing so, illustrate some of the mechanisms or approaches which might be useful in governing other aspects of data governance touching upon machine learning.

3.2. Line 2: Data from Models

Nature of Line 2 The relationship between data protection law and machine learning outlined in previous sections, in particular Chapter 2: *The Law of Machine Learning?* draws a clear line between the governance of the model, and the governance of data *before* turned into a model, and when the model is *applied* (to new personal data). In short, this classical view sees data protection law as already governing the collection and use of data in generating machine learning models, and, under certain limited conditions, the application of model results upon data subjects. For example, and as described above, (i) models cannot be trained from personal data without a specific lawful ground, such as consent, contract or legitimate interest; (ii) data subjects should be informed of the intention to train a model and (iii) usually maintain a right to object or withdraw consent; and (iv) in situations where models inform a significant, solely automated decision, individuals can appeal to the data controller for meaningful information about the logic of processing, or to have the decision taken manually reconsidered. The use of machine learning to turn 'normal' personal data into 'special category' personal data, such as race, political opinion or data concerning health, also requires establishing a lawful basis, which will usually be more stringent than

personal data in general.⁷²⁴ In certain situations, data controllers will be obliged to produce and update DPIAs concerning their use of machine learning technologies.

This entire approach relies on a clear line between models and personal data to exist. In this section, I argue this is not the case, and why this phenomenon might help us envisage new (although not necessarily recommended) ways of governing machine learning within a data protection framework.

3.2.1. Why Control Models?

If data protection already implicates machine learning in the above ways, does it matter if a model is personal data or not? In short—yes. The GDPR’s classic connection to machine learning is largely intermediated by the notion of personal data, governing systems only when such data are involved in training them or querying them to apply their results. Yet provisions would have a significantly different effect were models themselves to benefit from some of the status that personal data has. As it stands, there are no data protection rights nor obligations concerning models in the period *after* they have been (lawfully) built, but *before* any decisions have been taken about using them. As already described, the provisions that fall when such decisions are made are relatively limited in scope, applicability and utility.

Even the provisions that apply outside this period are relatively minor: those models being used as decision support, rather than as decision-making instruments, or those significantly affecting groups rather than individuals, are subject to few solid protections at all.⁷²⁵

Insofar as they *are* data, models would be considered ‘non-personal’ data,⁷²⁶ falling outside of data protection law,⁷²⁷ and potentially could be seen as subject of an intellectual property claim or trade secret⁷²⁸ or, feasibly, a database governed by the *sui generis* European database right.⁷²⁹

There is a reason to believe the level of control supplied by these flavours of legal status, which do not generally consider those affected by these systems downstream,

⁷²⁴ See further section 3.3.5.

⁷²⁵ See section 2.2.1.1 and 2.3.2.

⁷²⁶ This is a topic of recent regulation attempting to stop localisation laws around ‘non-personal data’, which defines it only as data ‘other than personal data’. As will be seen, the context specific nature of the definition of personal data makes this sort of definition very tricky, and the lines much more difficult to delineate than this new regulation seems to treat it as. See Regulation (EU) 2018/1807 of the European Parliament and the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union [2018] OJ L303/59 (Free Flow of Non-Personal Data Regulation) art 2(1).

⁷²⁷ GDPR, art 2(1) sets the scope of the law, and somewhat redundantly, GDPR, recital 26 emphasises that the ‘principles of data protection should therefore not apply to anonymous information’.

⁷²⁸ See generally Malgieri, ‘Trade Secrets v Personal Data: a possible solution for balancing rights’ (n 716).

⁷²⁹ Database Directive.

3. Data Protection's Lines, Blurred by Machine Learning

is insufficient. Individuals might want to control how they specifically are 'read' by machine-learned systems,⁷³⁰ particularly if individuals subscribe to the belief (often noted for its Germanic origin) in a right to informational self-determination (*informationelle Selbstbestimmung*).⁷³¹ As already noted, trained models can transform seemingly non-sensitive data, such as gait or social media use, into sensitive data, such as information on an individual's fitness or medical conditions.⁷³² In many fields, these models are far from user-independent, and so individuals adding their own granular training data to these models enables them to predict more accurately in the future. Take automated lip-reading systems—a topic elaborated further on later in this chapter.⁷³³ These transform videos of speech into approximated transcripts, and, being largely speaker dependent,⁷³⁴ require individual-specific training data to be effective. Once a model integrates an individual's data, it will predict their speech with significantly greater accuracy than that of others. Do individuals have a right to reverse this process, and have some agency over the model after it has been trained and potentially traded? Do they even have some right to use the model, or at least to know what it is used for, given their stake in training it?

In a similar vein, arguments have been forwarded that 'groups' of individuals deserve agency over their representation in models. Scholars concerned with 'categorical' or 'group' privacy see such groups as having collective rights to determine how their identities are constituted.⁷³⁵ Within this view, an ethical breach might be considered to have occurred 'when data or information is added to subject's identity without consent'.⁷³⁶ For example, given claims of correlations between smartphone-captured data and aspects of physical/mental health,⁷³⁷ individuals revealing a rare condition to a modeller might have unintentionally enabled such inferences to be successfully made for others, too. Similar arguments also exist in relation to data that connect individuals more deeply, such as genomic data, where sequencing the genome of one family member might reveal information about many.

⁷³⁰ Hildebrandt, *Smart technologies and the End(s) of Law* (n 68).

⁷³¹ See generally Gerrit Hornung and Christoph Schnabel, 'Data Protection in Germany I: The Population Census Decision and the Right to Informational Self-Determination' (2009) 25(1) *Computer Law & Security Review* 84 DOI: 10/d2zf3z.

⁷³² See eg section 1.5 and the Target war-story (section 1.4.3).

⁷³³ See below in section 3.3.

⁷³⁴ Helen L Bear, Stephen J Cox and Richard W Harvey, 'Speaker-independent machine lip-reading with speaker-dependent viseme classifiers' in *The 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing (FAAVSP-2015), Vienna, Austria, September 11-13, 2015* (2015).

⁷³⁵ Anton Vedder, 'KDD: The challenge to individualism' (1999) 1(4) *Ethics Inf. Technol.* 275; Linnet Taylor, Luciano Floridi and Bart van der Sloot, *Group Privacy* (Springer 2017).

⁷³⁶ Brent Mittelstadt, 'From Individual to Group Privacy in Big Data Analytics' (2017) 30(4) *Philos. Technol.* 475 DOI: 10/cwdg, 482.

⁷³⁷ See eg Farhan and others, 'Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data' in *2016 IEEE Wireless Health (WH)* (2016) DOI: 10/cwdh.

3.2.2. Models on the Move

These issues are of increasing importance given how data controllers increasingly refrain from trading data, as the ability to do this freely is heavily limited by data protection law, and instead look to trade or rent out models trained on it, as a way to pass on the value with fewer privacy and regulatory concerns. Many large firms already offer trained models for tasks including face recognition, emotion classification, nudity detection and offensive text identification. Two main business models underpin this practice.

The first is in the licensing of application programming interfaces (APIs) through ‘App Store’-like platforms. Firms earn royalties when their models are deployed. Microsoft’s Cortana Intelligence Gallery and Algorithmia’s Marketplace match API providers with potential customers. Google’s Cloud AutoML uses *transfer learning*,⁷³⁸ where insights from one modelling domain can be moved to another learned model, to allow firms to enrich Google’s generic models with their own specialised datasets. The trend towards augmentable, pre-trained ‘learnware’ appears to be a salient one.⁷³⁹

The second is the trading of packaged models. This might be preferable where APIs over the Internet are too sluggish, where queries are highly sensitive (eg medical records) or where transparency requirements require full access to model specifications. Apple’s recent Core ML library and Google’s Tensorflow Mobile are designed to run pre-trained models, some of which they provide, on portable devices.⁷⁴⁰

Model trading is an attractive prospect in relation to many aspects of privacy. It stands in stark contrast to the incumbent (and likely largely illegal) system of large-scale, indefinite data accumulation and retention by shadowy and distant data brokers—systems presenting significant difficulties for data subject comprehension and control.⁷⁴¹ While it is commonly claimed that the ‘free flow’ of data is of economic benefit, much of this benefit derives from the movement of mined insights rather than the transmission of individual records. Given that the push towards open data might, in at least some regards, be at tension with privacy,⁷⁴² model trading might serve as a

⁷³⁸ See generally Karl Weiss, Taghi M Khoshgoftaar and DingDing Wang, ‘A Survey of Transfer Learning’ (2016) 3(1) *Journal of Big Data* DOI: 10/gfkr2w.

⁷³⁹ Zhi-Hua Zhou, ‘Learnware: on the future of machine learning’ (2016) 10(4) *Front. Comput. Sci* 589 DOI: 10/cwdj.

⁷⁴⁰ See <https://developer.apple.com/documentation/coreml> and <https://www.tensorflow.org/lite/>.

⁷⁴¹ Max Van Kleek, Ilaria Liccardi, Reuben Binns, Jun Zhao, Daniel J Weitzner and Nigel Shadbolt, ‘Better the Devil You Know: Exposing the Data Sharing Practices of Smartphone Apps’ in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI’17)* (ACM 2017) DOI: 10/gfgq8q; Max Van Kleek, Reuben Binns, Jun Zhao, Adam Slack, Sauyon Lee, Dean Ottewell and Nigel Shadbolt, ‘X-Ray Refine: Supporting the exploration and refinement of information exposure resulting from smartphone apps’ in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI’18)* (ACM 2018) DOI: 10/cvcn; Binns, Lyngs, Van Kleek, Zhao, Libert and Shadbolt (n 568).

⁷⁴² David Banisar, *The right to information and privacy: balancing rights and managing conflicts* (World Bank

useful approach to balance trade-offs that emerge.

Furthermore, model trading might mitigate concerns around platform monopolies. Enabling users to deploy local personalisation tools might balance power relations against large firms hoarding personal data. This is the vision of personal data container proponents, who see distributed local processing, powered by decentralised, privacy-preserving analytical techniques—such as *secure multi-party computation* and *homomorphic encryption*—as enabling a shift of economic incentives and control from large players back to data subjects.⁷⁴³ Many processes that use highly sensitive or granular knowledge have been envisaged as better managed using edge computing, such as the delivery of advertisements or the personalization of news media,⁷⁴⁴ or discrimination auditing and ‘de-biasing’ of machine learning models,⁷⁴⁵ limiting the sensitive data leaving devices users directly control.

Yet such systems, no longer representing unique records which might render an individual identifiable, have not been considered as personal data, and thus have been considered excluded from the data protection regime. Next, I challenge this conventional understanding, and reflect upon the legal provisions this would trigger. Recent evidence, reviewed here, highlights that models themselves may leak data they were trained with—raising classic data confidentiality concerns. Data protection rights and obligations might then apply to models themselves. I outline the format model inversion attacks can take, why they might render models as personal data in the sense of European data protection law and what the consequences of this might be for data subjects and for data controllers.

3.2.3. Inverting Models

It has been demonstrated that machine learning models are vulnerable to a range of cybersecurity attacks that cause breaches of confidentiality. Confidentiality attacks leak information to those other than whom designers intended to view it. In the case of machine learning systems, there are different types of these attacks. The first concerns *model stealing*, eg where an attacker uses API access to replicate a model.⁷⁴⁶ Without a further confidentiality breach, this is primarily a concern for intellectual property rather than privacy, and of less concern here. A second attack class, *model*

2011).

⁷⁴³ Crabtree, Lodge, Colley, Greenhalgh, Mortier and Haddadi (n 679).

⁷⁴⁴ Jon Crowcroft, Anil Madhavapeddy, Malte Schwarzkopf, Theodore Hong and Richard Mortier, ‘Uncloaked Vision’ in Marcos K Aguilera, Haifeng Yu, Nitin H Vaidya, Vikram Srinivasan and Romit Roy Choudhury (eds), *Distributed Computing and Networking* (Springer 2011) DOI: 10/bj9n6t.

⁷⁴⁵ Kilbertus, Gascon, Kusner, Veale, Gummadi and Weller (n 263).

⁷⁴⁶ Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter and Thomas Ristenpart, ‘Stealing Machine Learning Models via Prediction APIs’ in *USENIX Security Symposium* (2016).

inversion, turns the journey from training data into a machine learned model from a one-way one to a two-way one, permitting the training data to be estimated with varying degrees of accuracy. A third attack class, *membership inference*, does not recover the training data, but instead recovers information about whether a particular individual was in the training set or whether they were not. Both model inversion and membership inference can be undertaken as a *black-box attack*, where the attack can be done with only query access (eg through the API business model above), or a *white box attack*, where an attacker requires full access to the model's structure and parameters.⁷⁴⁷

I will now formally describe both model inversion and membership inference attacks in a manner amenable to the discussion of personal data. A diagrammatic depiction of both accompanies the following description in Figure 3.2. The setup is as follows. A data controller holds a dataset of personal data $\mathbf{A} = a_{i,j} \in \mathbb{R}^{m \times n}$ where each $[a_{1,*}, a_{2,*}, \dots, a_{m,*}]$ within is a row of personal characteristics relating to one of the m data subjects in the set DS_1 , where $|DS_1| = m$, with each of the n variables indexed by j . They also have access to a model $M(\mathbf{B})$, which is a machine learned model trained on personal data $\mathbf{B} = b_{i,j} \in \mathbb{R}^{x \times y}$ where each $[b_{1,*}, b_{2,*}, \dots, b_{x,*}]$ within is a row of personal characteristics relating to one of x data subjects in the set DS_2 , where $|DS_2| = x$, and each one of the y variables a feature in the trained model. The data controller may have access to the model either directly (white-box) or via a query interface (black-box). I assume $DS_1 \cap DS_2 \neq \emptyset$: that is, some individuals are in both the training set and the additional dataset held. I refer to individuals in both DS_1 and DS_2 as set Z .

Under a model inversion attack, a data controller who does not initially have direct access to \mathbf{B} , but is given access to \mathbf{A} and $M(\mathbf{B})$, is able to recover some of the variables in training set \mathbf{B} , for those individuals in both the training set and the extra dataset \mathbf{A} . These variables connect to each other, such that the new personal dataset in question has all the variables of \mathbf{A} and some of \mathbf{B} . There may be error and inexactitude in the latter, but the data recovered from those in the training dataset will be more accurate than characteristics simply inferred from those that were not in the training dataset.

One of the earliest attacks resembling model inversion emerged from the field of *recommender systems*; a demonstration that collaborative filtering systems, where item recommendations are generated for a user based on behavioural patterns of other users, can end up revealing the consumption patterns of individual users.⁷⁴⁸ At the

⁷⁴⁷ Matt Fredrikson, Somesh Jha and Thomas Ristenpart, 'Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures' in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (2015) DOI: 10/cwmdm; Martin Abadi and others, 'On the Protection of Private Information in Machine Learning Systems: Two Recent Approaches' in *Proceedings of the 30th IEEE Computer Security Foundations Symposium, August 21-25, 2017, Santa Barbara, CA, USA* (2017).

⁷⁴⁸ Joseph A Calandrino, Ann Kilzer, Arvind Narayanan, Edward W Felten and Vitaly Shmatikov, "'You Might Also Like:' Privacy Risks of Collaborative Filtering' in *IEEE Symposium on Security and Privacy (SP)* (2011) DOI: 10/bfnjg3.

3. Data Protection's Lines, Blurred by Machine Learning

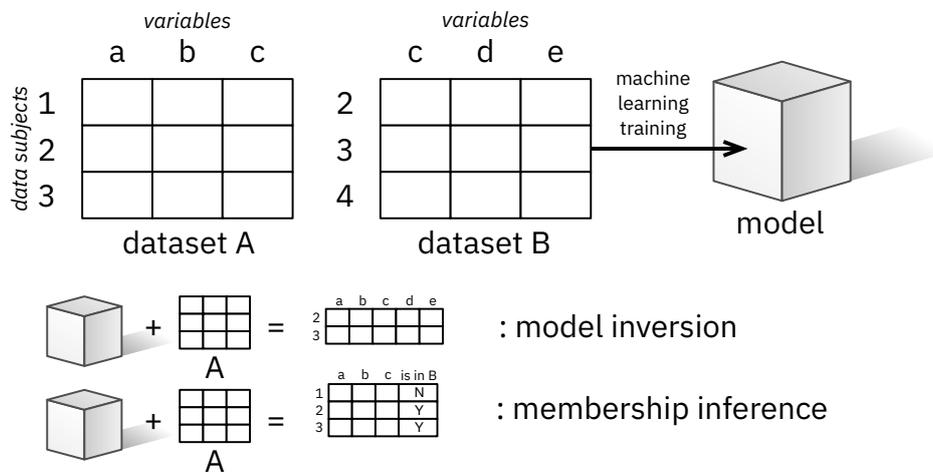


Figure 3.2.: Model inversion and membership inference attacks. Diagram by author.

time, a reader might have reasonably assumed such risks to be a quirk of the particular system and application area, but subsequent work suggests that this may be a more general problem potentially facing any kind of machine learning model that utilises personal data for training. Later work undertook attacks against several machine learning models to attempt to learn meaningful information about the training data. They were not concerned with privacy *per se*, but with trade secrets, seeking to uncover the ‘secret sauce’ of algorithms that might give them a commercial edge, such as whether speech recognition systems were trained on certain accents or not.⁷⁴⁹ Fredrikson et al. examine models designed to select correct medical doses for a widely used anticoagulant that interacts strongly with individual genetic markers. They show the possibility of reverse-engineering to reveal patients’ genetic markers with some demographic information about patients in the training data.⁷⁵⁰ Further work demonstrated both white- and black-box attacks re-identifying individuals from models trained on survey data with no false positives, and black-box attacks to reconstruct faces from facial recognition systems to the point where skilled crowd-workers

⁷⁴⁹ Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali and Giovanni Felici, ‘Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers’ (2015) 10(3) IJSN 137.

⁷⁵⁰ Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page and Thomas Ristenpart, ‘Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing’ in *USENIX Security Symposium* (2014). Note that some argue that this paper shows no privacy breach here beyond people finding machine learning ‘creepy’: see Frank McSherry, ‘Statistical inference considered harmful’ [2016] (<https://github.com/frankmcsberry/blog/blob/master/posts/2016-06-14.md>).

could use the photo to identify an individual from a lineup with 95% accuracy.⁷⁵¹

These attacks function in different ways, but broadly share a core structure. Initial demonstrations would focus on one or more input variables of interest they wanted to retrieve from the model, and use side-channel information that featured as model input (such as demographics of individuals) and permute the variable of interest, using the distribution of responses to understand which permutations were present in the training data. For variables with a huge feature space, such as images, the confidence level of the model returned upon prediction can be used to navigate the walk through the feature space of the input variable, using techniques including gradient descent.⁷⁵²

Other work connected to model inversion has indicated that only small changes to training algorithms lead to nearly indistinguishable models that are possible to exploit to leak large amounts of private data.⁷⁵³ or that systems exactly memorise specific private information in training sets, such as strings of sensitive data.⁷⁵⁴ Some model structures also require the training data in order to function, such as certain types of support vector machines, where the vectors that define the classification boundary, drawn from the training data, are bundled with the model used.

Membership inference attacks do not recover training data, but instead ascertain whether a given individuals' data was in a training set or not. Under a membership inference attack, the holder of \mathbf{A} and $M(\mathbf{A})$ does not recover any of the columns in \mathbf{B} , but can add an additional column to dataset \mathbf{A} representing whether or not a member of DS_1 is in the set Z : that is, whether or not they were also part of the training set participants DS_2 .

Shokri et al. demonstrate membership inference in a black box attack against a hospital discharge model.⁷⁵⁵ Connectedly, Pyrgelis et al. looked at membership inferences in location data, showing that it is comparatively easy to examine whether an individual with a certain movement pattern was used in the construction of an aggregate.⁷⁵⁶ They note this could be concerning if such aggregates were themselves made

⁷⁵¹ Fredrikson, Jha and Ristenpart (n 747).

⁷⁵² See generally *ibid*.

⁷⁵³ Congzheng Song, Thomas Ristenpart and Vitaly Shmatikov, 'Machine Learning Models that Remember Too Much' in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)* (2017) DOI: 10/cwdp.

⁷⁵⁴ Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson and Dawn Song, 'The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets' [2018] arXiv preprint (<https://arxiv.org/abs/1802.08232>).

⁷⁵⁵ Reza Shokri, Marco Stronati, Congzheng Song and Vitaly Shmatikov, 'Membership Inference Attacks Against Machine Learning Models' in *2017 IEEE Symposium on Security and Privacy (SP)* (2017) DOI: 10/cwdq.

⁷⁵⁶ Apostolos Pyrgelis, Carmela Troncoso and Emiliano De Cristofaro, 'Knock Knock, Who's There? Membership Inference on Aggregate Location Data' in *Proceedings of the 25th Network and Distributed System Security Symposium (NDSS 2018)* (2018) (<https://arxiv.org/abs/1708.06145>).

to understand a sensitive subgroup, such as individuals with dementia.

3.2.4. Models as Personal Data?

While these attacks are nascent, their potential is being increasingly demonstrated.⁷⁵⁷ The initial important question for data protection law is: to what extent would systems vulnerable to attacks like these be considered datasets of personal data, breaking the traditional line data protection draws between these concepts?

It is possible to make a compelling legal argument that a model inversion-vulnerable $M(\mathbf{B})$ would be personal data. A direct analogy can be made to personal data which has been 'pseudonymised'. The GDPR defines pseudonymisation as 'the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person'.⁷⁵⁸ Under the GDPR, pseudonymised data explicitly remains personal data. In the above setup, $M(\mathbf{B})$ represents the pseudonymised version of the training set \mathbf{B} , while \mathbf{A} represents the key by which such data can be partially reidentified. Where a single data controller is in possession of both \mathbf{A} and $M(\mathbf{B})$, and a model inversion attack is possible, it would appear by analogy that not only \mathbf{A} but also $M(\mathbf{B})$ should be considered personal data.

Of potentially greater interest however is the situation where a model $M(\mathbf{B})$ has been released, and so \mathbf{A} and $M(\mathbf{B})$ are held by different entities. There is legal precedent for the model $M(\mathbf{B})$ to be considered personal data in this case too.

Personal data means any information relating to an identified or identifiable natural person.⁷⁵⁹ In many cases, individuals can be identified in supposedly anonymous datasets using external sources of data,⁷⁶⁰ which creates a great deal of uncertainty around which datasets are *not* personal data.⁷⁶¹ The exact ease of de-anonymisation is however highly context specific, and some researchers have pushed back on the alluring rhetoric of 'surprising ease' of reidentification and 'broken promises' of anonymisation to argue that it is often harder and more context-specific that high profile

⁷⁵⁷ I remain cautious about overstating their practical efficacy, which does remain unclear—I not aware of any documented attacks 'in-the-wild'.

⁷⁵⁸ GDPR, art 4(5).

⁷⁵⁹ GDPR, art 4(1).

⁷⁶⁰ See eg Paul Ohm, 'Broken promises of privacy: Responding to the surprising failure of anonymization' (2009) 57 UCLA L. Rev. 1701; Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh and Alex Pentland, 'Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata' (2015) 347(6221) Science 536 DOI: 10/zt7; Montjoye, Hidalgo, Verleysen and Blondel (n 616).

⁷⁶¹ Elliot, Mackey, O'Hara and Tudor (n 563); Purtova (n 364).

studies have led the field to believe.⁷⁶²

In the recitals, the GDPR provides an aide to navigating this problem, with a test of reasonable likelihood of reidentification.⁷⁶³

To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.

When asked in recent years, the CJEU has set a low bar and wide scope for what is considered personal data. In *Breyer*, the CJEU clarified the reasonable likelihood test in the DPD that now forms part of the recital above.⁷⁶⁴ In particular, the CJEU took a wide view of identifiability, clarifying that where additional data is required to reidentify a dataset, that data need not all be in the hands of a single entity to be considered personal data.⁷⁶⁵ In effect, this illustrates that in the case of either attack M(**B**) might be considered personal data if dataset **A** is held by another entity.

Furthermore, while the personal data in model inversion attacks is quite easily construed as personal data, insofar as it resembles a training set or can be used to identify individuals, it might initially appear less clear that data obtained from membership inference—whether an individual was in a training dataset—is personal data. Personal data does not specify particular sensitive or private categories in the definition, instead using the phrase ‘any information’. In *Nowak*, the CJEU engaged with the meaning of this phrase, noting that it is ‘not restricted to information that is sensitive or private, but potentially encompasses all kinds of information, not only objective but also subjective’.⁷⁶⁶ Similarly, and following the A29WP, the Court understood that this ‘any information’ can ‘relate to’ an individual in many ways: by content, by purpose and by effect.⁷⁶⁷ Seen through this lens, information on an individuals’ membership of a training set would indeed fall within the scope of personal data, regardless of how trivial or mundane it might be to the individual it concerns.

⁷⁶² See eg Mark Elliot and others, ‘Functional Anonymisation: Personal Data and the Data Environment’ (2018) 34(2) Computer Law & Security Review 204 DOI: 10/gdhs4w.

⁷⁶³ GDPR, recital 26.

⁷⁶⁴ Case C-582/14 *Patrick Breyer v Bundesrepublik Deutschland* ECLI:EU:C:2016:779.

⁷⁶⁵ *ibid* ¶ 43 (‘it is not required that all the information enabling the identification of the data subject must be in the hands of one person’).

⁷⁶⁶ *Nowak* (n 414) ¶ 34.

⁷⁶⁷ *ibid* ¶ 35.

3. Data Protection's Lines, Blurred by Machine Learning

There does exist an argument in the scholarship which argues that the CJEU's approach to the scope of personal data fuels undesirable data protection maximalism. Purtova considers a situation where common environmental factors which do not identify an individual by themselves, such as the weather, are used in a smart city context as an input to a behavioural profiling system.⁷⁶⁸ In this case, she argues, it is plausible that the weather data would be personal data by means of purpose and effect. In the context of this paper, her argument might plausibly be extended to claim that the weights, and perhaps the structure, of machine learning models relate to an individual by means of impact, and by virtue of this are personal data. While I acknowledge this *reductio ad absurdum* argument concerning the current scope of personal data, and the consequences for it as making the law impracticably broad, this argument does not lean in this direction. I do not aim to critique this analysis, nor support, oppose or resolve the dilemma raised by Purtova; but merely to note that the argument made here—that inverted models might fall under the definition of personal data—does not depend on the kind of expansive definition that might give rise to such absurdities. Thus, even if the definition were to be somehow tightened in scope,⁷⁶⁹ the argument above concerning inverted models would still likely stand.

It should also be noted that in some cases, models are designed to encode data which might be personal. For example, natural language processing applications include question–answer models trained on large corpora of text, which can be queried for biographic data about individuals despite this being opaquely coded in the weights of a model. In these special cases, a model inversion attack is not even required to make an argument that these systems should and could be considered as personal data.⁷⁷⁰

In sum, model inversion and membership inference attacks, where possible, do risk models being considered as personal data even without resorting to a the type of maximalist reading of data protection law that Purtova highlights is possible. The line is blurred—not everywhere or all the time, but the potential remains. A clear question then follows—what are the practical consequences of this, and are they of interest to those hoping to better control these systems and hold them to account? It is to these consequences I now turn—firstly, the implications with data subjects at their core, and secondly, the implications for data controllers.

⁷⁶⁸ Purtova (n 364).

⁷⁶⁹ As the Court did between Case C-141/12 *YS v Minister voor Immigratie, Integratie en Asiel and Minister voor Immigratie, Integratie en Asiel v M and S* ECLI:EU:C:2014:2081 and *Breyer* (n 764). See further Purtova (n 364).

⁷⁷⁰ See eg Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever, 'Language Models Are Unsupervised Multitask Learners' [2019] OpenAI Working Paper.

3.2.5. Implications for Data Subjects

Data subject rights, as already discussed, are core to data protection. Here, I consider three tasks data subjects might want to achieve in relation to models trained with their data, and how model inversion might interact with them: where a data subject wishes to access models and to know where they have originated from and to whom they are being traded or transmitted; where a data subject wishes to erase herself from a trained model; and where a data subject wishes such a model not to be used in the future.

3.2.5.1. Information Rights

As seen in previous sections, GDPR contains information provisions which are triggered on request, such as the right of access and the right of portability, as well as provisions where the data controller must provide information to the data subject without being solicited to do so.⁷⁷¹

In practice, these rights are most well-known for the requirement, carried through from the DPD, to provide all the information held about an individual upon request. It seems unlikely that such a request would allow an individual to request an entire model on the basis that it comprised in part of their data as, except in the rare case where only a single individual's data was used for the model building,⁷⁷² as to do so might comprise of the privacy of others in the training set (for the very same reason as the model would qualify as personal data). To provide it in that way would be a breach of data protection's security principle. It also seems unlikely that they would be successful with much ease at requesting a copy of the data retrieved from a model, as, just as described above, it is unlikely to include a name or identifier, and to make names and identifiers difficult to access is a common practice by data controller which has the effect of rendering personal data difficult to exercise rights over.⁷⁷³

Particularly given the increasing interest in model trading described above, it is more interesting to consider other aspects of these information rights. In particular these rights might enable better tracking of the provenance of data, analysis and decision-making systems, something which has received increasing attention of late.⁷⁷⁴ In effect, these rights help track the origin and destination of trained models.

⁷⁷¹ See section 2.2.2.1.

⁷⁷² Such a case might include, for example, a tailored voice recognition system or a learned smart home personalisation model. Portability of these models to another controller might be useful, although it would introduce a likely infeasible compatibility demand on the receiving provider, who would likely find it easier simply to take the raw data and retrain a model themselves.

⁷⁷³ See section 3.1, p. 139.

⁷⁷⁴ Singh, Cobbe and Norval (n 264).

3. Data Protection's Lines, Blurred by Machine Learning

The GDPR does have two core types of requirement which respectively require controllers with data to provide information about from whom specifically personal data came, and to whom, more generally, it is going.

At the point of collection,⁷⁷⁵ and upon request,⁷⁷⁶ the data subject should be provided with 'the recipients or categories of recipients of the personal data, if any'. While interpretations from DPAs have indicated that providing potentially broad and unhelpful 'categories' alone is in line with the law,⁷⁷⁷ recent guidance from the A29WP notes that only providing the 'categories' will require a justification of why this is in line with the principle of fairness in data protection, and if it is done, such categories must 'be as specific as possible by indicating the type of recipient (ie by reference to the activities it carries out), the industry, sector and sub-sector and the location of the recipients.'⁷⁷⁸

Where one controller receives personal data from another—as is the case with model trading if we conceive of models that way, there is an alternative approach afforded by Article 14(2)(f). This provision states that 'the controller shall provide the data subject with the following information necessary to ensure fair and transparent processing in respect of the data subject [...] from which source the personal data originate, and if applicable, whether it came from publicly accessible sources'.

The only applicable exemption here is where 'the provision of such information proves impossible or would involve a disproportionate effort'.⁷⁷⁹ It is highly likely that a controller faced with the situation of model-as-personal-data would try to utilise this exemption, particularly as they are unlikely to have contact information for the data subjects in the model, and the GDPR notes they are not obliged to hold such additional data for the sole purposes of complying with the Regulation.⁷⁸⁰ This is however not the end of the story, because Article 14(5)(b) goes on to note that where disproportionate effort applies, '[in] such cases the controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available.' The one safeguard that is explicitly listed would seem to imply that the data controller might then be obliged to make publicly available—perhaps through their website—their receipt of a model and the sources from which the personal data originate.

In effect, this would mean that organisations receiving invertible or potentially in-

⁷⁷⁵ GDPR, art 13(1)(e).

⁷⁷⁶ GDPR, art 15(1)(c).

⁷⁷⁷ Antonella Galetta and Paul de Hert, 'Exercising Access Rights in Belgium' in Clive Norris, Paul de Hert, Xavier L'Hoiry and Antonella Galetta (eds), *The Unaccountable State of Surveillance: Exercising Access Rights in Europe* (Springer 2017).

⁷⁷⁸ Article 29 Data Protection Working Party, *Transparency Guidelines* (n 701).

⁷⁷⁹ GDPR, art 14(5)(b).

⁷⁸⁰ GDPR, art 11(1).

vertible models would have to publish where they came from. To do so would be a huge asset to those mapping these flows and looking to understand the data economy and algorithmic accountability within it, particularly if such notices were machine readable in form.

3.2.5.2. Erasure Rights

The famous right to be erasure, known by some by the politicised moniker of the *right to be forgotten*, is a qualified right of a data subject to ‘the erasure of personal data concerning him or her’.⁷⁸¹ Core reasons a data subject might want to erase herself from a model overlap with the general reasons for model control presented above⁷⁸²—to erase insights about her she might dislike; to erase unwanted insights about a group she identifies as part of; or to erase insights which might lead to data breaches.

There are two main ways to erase data from a trained model.

Firstly, a model can be trained based upon an amended training dataset, with the person requesting erasure omitted. The computational intensity of much machine learning training, even on the world’s most powerful computational infrastructures, does however mean such training is far from cost-free or instantaneous. Retraining brings significant energy costs—data centres consume between 1.1–1.5% of global power consumption,⁷⁸³ as well as time and labour costs. Furthermore, and as noted previously, insofar as firms have deleted explicit identifiers from the training set, practically locating the individual to omit may be challenging.⁷⁸⁴

Secondly, the model itself can be amended after training. This is not easy, and currently rarely possible in modern systems. Approaches for quick and easy ‘machine unlearning’ are only beginning to be proposed and are still largely unexplored, let alone at a stage ready for deployment.⁷⁸⁵ Currently proposed methods cannot be retrofitted onto existing systems, and would require entire model pipelines to be re-conceived, with unclear effects. Consequently, large-scale organisations such as Google do not appear to remove links from their trained search model, but filter (‘delist’) results between model output and search result delivery.⁷⁸⁶

⁷⁸¹ GDPR, art 17.

⁷⁸² See section 3.2.1, p. 173.

⁷⁸³ Toni Mastelic, Ariel Oleksiak, Holger Claussen, Ivona Brandic, Jean-Marc Pierson and Athanasios V Vasilakos, ‘Cloud Computing: Survey on Energy Efficiency’ (2015) 47(2) ACM CSUR 33 DOI: 10/cwd6.

⁷⁸⁴ See section 3.1, p. 139.

⁷⁸⁵ Yinzhi Cao and Junfeng Yang, ‘Towards Making Systems Forget with Machine Unlearning’ in *2015 IEEE Symposium on Security and Privacy* (2015) DOI: 10/cwd7; Debjanee Barua, ‘A time to remember, a time to forget: User controlled, Scalable, Life long user modelling’ (Doctoral dissertation, The University of Sydney 2016).

⁷⁸⁶ Google, *Transparency Report* (<https://perma.cc/8DE4-AXBW>).

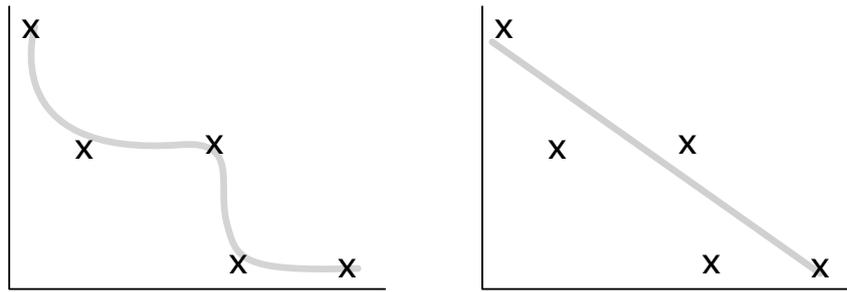


Figure 3.3.: A simple example of overfitting. An arguably overfitted regression is on the left, whilst a more intuitively well-fitted regression is on the right. Which is more appropriately fitted in practice however would depend on the 'real' distribution of future data, and which line it agreed with most.

Yet while it might be possible for a single individual to remove a pattern from a system trained on a great deal of data from a few individuals, where the number of observations in a dataset is similar to the number of data subjects, it would appear difficult for an individual to remove such patterns alone (except from poorly trained models). Models basing patterns on single data records are generally considered to have been *overfitted*—memorising the training data rather than finding generalisable patterns.⁷⁸⁷ An illustration of overfitting can be found in Figure 3.3. It could be argued that if an erasure right mattered, the machine learning system was initially trained poorly to begin with. Then again, model inversion attacks seem especially amenable to overfitted models that have 'remembered' data in such granular detail, and thus it may be especially important or useful to exercise a right to erasure over overfitted models rather than models in general. Furthermore, overfitting is not a simple trade-off or parameter. A model might be overfitted on certain parts of the phenomenon, and more 'appropriately' trained on other aspects. These issues serve to complicate simple analysis.

Such a right is arguably more useful when used collectively. For example, users of a certain demographic group, sharing a certain medical condition, living in a certain neighbourhood, might see an interest in having their entire group erased from the model, and might seek to either completely scrub a model of data relating to them, or just reduce it to a level where it functions badly. Whether this is possible will depend both on the phenomena at hand, and the level of coordination between the data

⁷⁸⁷ There are times when one record can make a difference: 'one-shot learning' in image recognition, for example. As children, humans are very good at seeing an object once, classifying it, and then correctly identifying a second, different object of the same classification if presented. Standard machine learning systems perform very poorly at this task, requiring many examples of a single class in order to become good at recognising it.

subjects—a classic collective action problem.⁷⁸⁸ Co-ordination in the form of online petitions and crowdfunding make this an interesting avenue to explore for the right of erasure, and one that is already being probed in relation to the right of access.⁷⁸⁹

3.2.5.3. Restriction and Objection Rights

A set of important rights that have received little media attention essentially surround the rights of individuals to say no to types of processing which they did not consent to. If models were personal data, the consequences of this are quite difficult to say. The right to object⁷⁹⁰ allows an individual to object to processing based on the legitimate interests of the data controller or public interest (largely public sector processing), as long as the data controller cannot demonstrate a ‘compelling’ legitimate interest of their own—something which seems a high bar indeed, and seemingly unlikely to be met unless some social benefit is present, rather than just controllers’ economic incentive.

The right to restrict processing⁷⁹¹ has a wider range of provisions, broadly giving the data subject rights to stop processing of their data until, for example, a balancing test relating to the right to object can be carried out, or until the accuracy of the data is ascertained and rectified. These are time-limited, and while there is no balancing test to use them (and therefore they could be used quite disruptively), they are generally considered lesser.

What is it to object to the use of a model? To query a model might be considered analogous to consulting the personal dataset within. Consultation is one of the many actions that explicitly comprise processing in data protection.⁷⁹² Yet because the data is not organised by record when it is in the form of the model, querying a model seems more like querying the entire dataset than querying a single record. Does objection give every individual a veto right over the consultation of the model in its entirety, for any lawful purpose? This would seem highly problematic, and likely disproportionate, but it is a possible reading of this provision, and one which demonstrates the difficulties seeing models as personal data would engender.

These difficulties stem from the fact that personal data in the GDPR is a wider concept than simply data which are organised by row and by column.⁷⁹³ For example,

⁷⁸⁸ On the difficulties of acting collectively, see generally Mancur Olson, *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard University Press 1965).

⁷⁸⁹ Mahieu, Asghari and van Eeten (n 493).

⁷⁹⁰ GDPR, art 21.

⁷⁹¹ GDPR, art 18.

⁷⁹² GDPR, art 4(2).

⁷⁹³ Note here that the slightly confusing definition the material scope of the Regulation laid out in GDPR, art 2(1) is that '[t]his Regulation applies to the processing of personal data wholly or partly by automated

3. Data Protection's Lines, Blurred by Machine Learning

data which are encrypted remain personal data due to the *potentiality* of it being decrypted. If an encrypted laptop containing highly sensitive data is left on a train, it may still count as a data breach even though the files within are not cleartext. Presumably, if such a system still consists of personal data, then consultation is still processing it. Decrypting a file of personal data would likely be considered processing it, even though it is not an activity carried out row-by-row, individual-by-individual. Yet objection and restriction rights more-or-less assume that all purposes are carried out on an individual-by-individual basis. How this uneven scaling plays with the conceptual form these rights take is therefore hazy indeed.

A more sensible approach from a user perspective to achieve their goals would often be to restrict the processing of that model in relation to an individual decision. This is already possible by using normal data protection rights to object or restrict the personal data constituting the *query* being used for prediction in a specific case. It may also be possible in certain cases using right not to be subject to solely automated decision-making⁷⁹⁴ which, unusually for data protection, targets decisions rather than data processing. Yet the strange interplay between objection and models made of mixed-up personal data could potentially be a place of tension were individuals to try to test and enforce these rights.

3.2.6. Implications for Data Controllers

Data controllers must consider both the specific rights and obligations they are subject to, as well as adherence to the overarching principles of data protection. I highlight two relevant areas here in the context of models as personal data: the security principle and the storage limitation principle.

3.2.6.1. Security Principle

As an overarching concern and obligation, data controllers need to consider whether their system leaks personal data in the first place. It seems unlikely that a modeller would wish to establish a legal basis for onward transfer of a model they have trained, something which would be especially onerous if that model is being transmitted outside of the EU or a country deemed 'adequate' with EU data protection law. If they do transfer such a model without a legal basis to do so, and such a model is inverted, it

means and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system.' Two readings of this are possible: that all data, automated or not, must fulfil the filing system criterion, or that only data being processed 'other by automated means' must do so. Grammatically, it seems that the latter reading makes more sense: indeed, the 'filing system' notion was designed to bring large paper-based records into the legal regime.

⁷⁹⁴ GDPR, art 22; see further section 2.2.1.

would likely be considered both a data breach and a violation of the security principle more generally. This is a relatively shocking conclusion, and endangers many of the positive aspects of model trading, particularly that large datasets are not being transmitted constantly, which this paradigm promises.

Further reason to ensure that models are not invertible comes from Article 25(1) of the GDPR, which introduces a qualified obligation to implement technical and organisational measures designed to implement data protection principles. In combination, these aspects require us to consider how models can be made which are resilient to the attacks described above.

Thankfully, security researchers try to secure systems as well as try to break them. The most common defence that can be levied against model inversion attacks is differential privacy:⁷⁹⁵ a defence discussed in most of the papers above describing the threats. Differential privacy is a formalised notion of privacy as information disclosure.⁷⁹⁶ At the core of differential privacy is the notion of adding precise forms of random noise to queries such that that for every individual in a dataset, removing them from that dataset will not noticeably change the results of that query.⁷⁹⁷ Differential privacy guarantees essentially limit what can be learned about any individual in the dataset to that which could be learned about them from everybody else's data alone. Differentially private approaches to machine learning work in a similar way: models lacking data relating to a particular individual function extremely similarly to those containing it.⁷⁹⁸ This provides intrinsic protection against a great deal of model inversion attacks.

Theoretically, such learning algorithms are just as powerful as non-differentially private ones.⁷⁹⁹ Since algorithms are supposed to find generalisable patterns, not memorise and regurgitate specific records, this makes intuitive sense (discussed fur-

⁷⁹⁵ Cynthia Dwork, 'Differential Privacy: A Survey of Results' in Manindra Agrawal, Dingzhu Du, Zhenhua Duan and Angsheng Li (eds), *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings* (Springer 2008) DOI: 10.1007/978-3-540-79228-4_1.

⁷⁹⁶ Differential privacy was popularised by Cynthia Dwork and her laboratory. See Cynthia Dwork, Frank McSherry, Kobbi Nissim and Adam Smith, 'Calibrating Noise to Sensitivity in Private Data Analysis' in Shai Halevi and Tal Rabin (eds), *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings* (Springer 2006). A useful lay introduction is Kobbi Nissim and others, *Differential Privacy: A Primer for a Non-technical Audience* (A product of the "Bridging Privacy Definitions" working group, part of the Privacy Tools for Sharing Research Data project at Harvard University 2017).

⁷⁹⁷ Pure differential privacy would mean that the analysis is *exactly* the same; in practice there is a small margin of flexibility allowed.

⁷⁹⁸ Zhanglong Ji, Zachary C Lipton and Charles Elkan, 'Differential Privacy and Machine Learning: a Survey and Review' [2014] arXiv preprint (<https://arxiv.org/abs/1412.7584>).

⁷⁹⁹ S Kasiviswanathan, H Lee, K Nissim, S Raskhodnikova and A Smith, 'What Can We Learn Privately?' (2011) 40(3) SIAM J. Comput. 793.

3. Data Protection's Lines, Blurred by Machine Learning

ther below). Yet despite the growing interest in differential privacy,⁸⁰⁰ the real challenge comes with deployment. The tools available today can be computationally expensive to deploy⁸⁰¹ as well as easily undermined with even small or arcane software errors.⁸⁰² Only a few large and powerful companies have demonstrated an ability to deploy them, and only then for very limited purposes.⁸⁰³ Furthermore, differential privacy works well at protecting disclosure in contexts where every individual has the same weight, such as in a count, but poorly in situations where there are extreme outliers. In a dataset of wealth containing Bill Gates, as the amount of noise needed to be added reduces the query results to absurdity.⁸⁰⁴ In cases where outliers might be vulnerable as well as powerful this is problematic: we might suspect it is they who need data protection the most.⁸⁰⁵

A second linked line of defence, albeit without the guarantees differential privacy provides, is to attempt to make models that do not 'overfit' the data. Where they do, they are confusing the 'noise' in the dataset for the generalisable 'signal' that helps prediction on unseen cases: memorising rather than doing anything close to learning. Avoiding overfitting is important—and in this way, data protection by design might legally oblige the training of methodologically sound models—but avoiding it is not enough to guarantee a model will not be vulnerable to model inversion. In some cases, such attacks have been shown to succeed in part even in the absence of overfitting.⁸⁰⁶

⁸⁰⁰ Andrew Orłowski, 'Apple pollutes data about you to protect your privacy. But it might not be enough' [2016] *The Register* (<https://perma.cc/98SX-CTTR>); Cory Doctorow, 'Data protection in the EU: the certainty of uncertainty' [2013] *The Guardian* (<https://perma.cc/DFY4-9SNC>); Tim Bradshaw, 'Apple plays catch-up with iMessage emojis' [2016] *Financial Times* (<https://perma.cc/LT9T-6FV9>).

⁸⁰¹ Kamalika Chaudhuri, Claire Monteleoni and Anand D Sarwate, 'Differentially private empirical risk minimization' (2011) 12 *J. Mach. Learn. Res.* 1069.

⁸⁰² Ilya Mironov, 'On Significance of the Least Significant Bits for Differential Privacy' in *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS'12)* (New York, NY, USA, 2012) DOI: 10/cwjw.

⁸⁰³ See eg Bolin Ding, Janardhan Kulkarni and Sergey Yekhanin, 'Collecting Telemetry Data Privately' in I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett (eds), *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc 2017) on Microsoft's deployments; Differential Privacy Team, Apple, 'Learning with Privacy at Scale' (2017) 1(8) *Apple Machine Learning Journal* (<https://perma.cc/T2RM-B27X>), on Apple's deployments; Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang and Xiaofeng Wang, 'Privacy Loss in Apple's Implementation of Differential Privacy on macOS 10.12' [2017] arXiv preprint (<https://arxiv.org/abs/1709.02753>), a criticism of Apple's deployments; Úlfar Erlingsson, Vasyl Pihur and Aleksandra Korolova, 'Rappor: Randomized aggregatable privacy-preserving ordinal response' in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (2014) DOI: 10/cwj2 on Google's deployments.

⁸⁰⁴ Krishnamurty Muralidhar and Rathindra Sarathy, 'Does Differential Privacy Protect Terry Gross' Privacy?' in Josep Domingo-Ferrer and Emmanouil Magkos (eds), *Proceedings of Privacy in Statistical Databases (PSD 2010)* (Springer 2010) DOI: 10/cr2zq7; Jane Bambauer, Krishnamurty Muralidhar and Rathindra Sarathy, 'Fool's Gold: An Illustrated Critique of Differential Privacy' (2013) 16 *Vand. J. Ent. & Tech. L.* 701.

⁸⁰⁵ Edwards and Veale, 'Slave to the Algorithm?' (n 79) 61.

⁸⁰⁶ Samuel Yeom, Irene Giacomelli, Matt Fredrikson and Somesh Jha, 'Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting' [2018] *IEEE Computer Security Foundations Symposium (CSF 2018)*.

3.2.6.2. Storage Limitation

Relatedly, another data protection principle, that of ‘storage limitation’, applies in this case. Storage limitation means that data should be kept for no longer than it is necessary for the purposes for which it is processed. As training data may need to be discarded as time goes on to meet this obligation, so might models. Some techniques from the field of concept drift adaptation could be useful here. This is the domain of research which looks at how to understand and model changing phenomena in machine learning systems. For example, a variety of methods exist to limit the use of older data in machine learning systems by both managing data used in machine learning and gradually forgetting data within a model.⁸⁰⁷ These methods are primarily used to better model phenomena today, particularly where old correlations may now be irrelevant, however similar efforts are likely to be required for any systems for which model inversion is not a readily remediable vulnerability.

3.2.7. Interim discussion

In this section, I have outlined how recent confidentiality attacks upon machine learning systems, including model inversion and membership inference, interplay with data protection law. Where models are vulnerable to such attacks, they gain an additional dimension—not only an analytic product potentially protected by intellectual property rights, but also a set of personal data, conceptually close to the idea of ‘pseudonymisation’ in the GDPR. In doing so, they blur a critical line that underpins much analysis of machine learning both to date and within this work. I illustrated a selection of consequences flowing from this new classification that are relevant to discussions of the governance of machine learned models which are being traded and transferred between data controllers. These include those of direct utility to data subjects, such as information, erasure and objection rights, and overarching obligations relevant to data controllers, such as security and storage limitation provisions. Seeing models as personal data could serve to re-balance or at least disrupt the power relations between those holding models and those whose data is used to train them.

There is however reason for caution amidst the promise. Whilst potentially enabling useful provisions, requiring what is essentially a security vulnerability in order to trigger rights and obligations is disconnected and arbitrary. Models not amenable to model inversion might still be models individuals wish to know the origin or destination of or wish to have themselves or their group erased from. I suspect many situations of problematic profiling will not be vulnerable to model inversion, and there-

⁸⁰⁷ Gama, Žliobaitė, Bifet, Pechenizkiy and Bouchachia (n 366).

fore not governable with this approach. Furthermore, where some model inversion is possible, technical challenges will make it difficult for data subjects and regulators to prove the leakiness. In many cases, while a companion dataset that may enable such an attack might exist, potentially on shady markets, it will not be held by the data subject, auditor nor the regulator, and thus the risk will be difficult to assess even if full model access is provided. The setup appears to further burden every stakeholder apart from the model, and it remains questionable whether downstream governance provisions in general, such as rights after model training, are the best way to deal with algorithmic harms at all.⁸⁰⁸

On one hand, the GDPR has created a set of principles that are desirable and robust for many purposes. On the other hand, it is showing that seeing data protection as an omnibus governing regime for all data-driven issues is misguided. In this chapter, I illustrate this by considering model inversion as a regulatory experiment—one that is potentially realistic, although far from advisable as a foundation for future algorithmic governance. I argue it provides means to probe whether or not it is appropriate these rights and obligations extend to analytic products themselves, and what the consequences of this development might be. Some consequences, such as the mapping of the provenance of trained models, seem potentially useful in oversight of increasingly influential systems. Others, such as the right to object and restrict processing, seem at tension with the very notion of model inversion. As the domain of personal data expands, it is important to recognise that while its scope is wide, its scope of effective and enforceable governance might not be. Pushing the boundaries of data protection law in light of new technologies serves as a useful and clarifying force in the continuous task of the coming decades in better applying existing law, and developing regulatory regimes around societal issues current regimes fail to deal with in practice.

3.3. Line 3: Data from Sensitive Data

The Nature of Line 3 Data protection imposes quite different requirements on ‘ordinary’ personal data from ‘special category’ personal data. This line goes back a long way: in pan-European terms, since the 1995 DPD. Yet machine learning in particular, transforming data into new forms and insights with comparative ease and limited effort or resource, blurs this regulatory line heavily. The ‘special categories’ of data in the GDPR are not the only form of sensitive data to become subject to this blurring, however. In this section, I consider an application—automated lipreading—that blurs *many* forms of sensitive data with data that there previously was a relatively well ac-

⁸⁰⁸ See section 2.3.1 for a discussion of the *transparency fallacy*.

cepted social contract for collecting (at least in the UK)—CCTV. In doing so, I will illustrate the challenges to the line between sensitive and non-sensitive data that come more widely from machine learning, as unlike the tabular or filing-system metaphor of data protection dealing in ‘categories’, machine learning is particularly adept at reclaiming insights from data such as images or text, which resist having such simple or clear lines drawn through them.

3.3.1. Automated Lipreading

The private sector’s increased use of inference and ambient tracking technologies, such as those based on WiFi and Bluetooth, has not slipped unnoticed past data protection authorities in the EU.⁸⁰⁹ The transformation of the nature of more familiar tracking modalities (such as video surveillance) thanks to algorithmic systems and particularly machine learning has, however, received less examination and scrutiny. In this section, I examine one emergent technology for mining further data from video—automated lipreading systems (ALRSs). Such technologies, primarily today using machine learning, can reconstruct speech from videos of faces and lips. Their use, particularly in relation to video data captured in public spaces, is likely to prove controversial.

Firstly, I will compare ALRSs to the related, existing tools such as face recognition⁸¹⁰ and affective computing (ie emotion detection), and discuss the likely short- and medium-term trends in research and application, justifying why they are likely to have a serious impact on personal privacy. Researchers currently present ALRSs as improving user experience and accessibility, but on examination it is argued that a significant potential application of these tools will lie in their ability to reconstruct the conversations of individuals and crowds, leading to a further refinement of existing profiling activity. Less dismally, lipreading technologies may also bring opportunities for accessibility that could change lives for the better. Yet there remains a near-vacuum of commentary⁸¹¹ concerning broader social aspects of these systems,

⁸⁰⁹ See eg Information Commissioner’s Office, *Wi-fi location analytics* (n 613); College bescherming persoonsgegevens (n 611); Mavroudis and Veale (n 611).

⁸¹⁰ See generally International Working Group on Data Protection in Telecommunications, *Working Paper on Intelligent Video Analytics 58th Meeting, 13-14 October 2015, Berlin (Germany)- 675.51.11* (Datenschutz Berlin 2015) (https://www.datenschutz-berlin.de/pdf/publikationen/working-paper/2015/14102015_en_2.pdf).

⁸¹¹ There appears to be little reflection in either research or the news media on privacy aspects of these technological developments. Two of the only pieces located which discuss this aspect in brief appear to be Anonymous, ‘The Challenges and Threats of Automated Lip Reading’ (*MIT Technology Review*, 11th September 2014) (<https://perma.cc/3MEX-FUL9>); Peter Swindon, ‘Lip-Reading CCTV Set to Capture Shoppers’ Private Comments for Big Companies’ (*Sunday Herald (Scotland)*, 17th August 2017) (<http://perma.cc/U5KH-XG9A>).

which is concerning given how even poorly performing systems could have a concerning broader effect.

I then discuss how the machine learning-powered ability of ALRSs to generate new personal data from video collected at a distance, often covertly and beyond the reasonable expectation of data subjects, may be more intrusive and less well-regulated than the basic paradigm of CCTV data collection. The current regulation of lip-reading is examined, both by comparison with CCTV and audio surveillance, in particular by reference to European data protection law.

Finally, I consider the interplay between ALRSs and DPbD, highlighting in particular some of the ongoing technical challenges that might cause issues even if considering advanced methods of using these technologies in low-risk and minimally invasive ways. I consider how data minimisation, purpose limitation, and the right to objection might play out in relation to these technologies, and end with some general conclusions around how the direction of these technologies might require us to reconsider the governance of video data in light of consequential machine learning.

3.3.2. Trajectories

Similar to many technologies suddenly entering public view, research into automated lip-reading is far from new. The first automatic lip-reading technology was demonstrated in 1984⁸¹², and several hundred papers have been published in the years following.⁸¹³ Early researchers felt that 'visual speech decoding' would be a relatively easy task, especially given successes seen in audio-based decoding (ie speech recognition).⁸¹⁴ Lipreading turned out to be significantly more challenging than envisioned.⁸¹⁵ Indeed, the task commonly challenges humans. While we excel at many of same speech tasks where we have seen machine learning breakthroughs in recent years, such as recognition (normal listening) or synthesis (speaking or doing impressions), even experts struggle with lip-reading. Studies attempting to estimate average human accuracy of lip-reading have placed it at around 12.4% on visual-only activit-

⁸¹² Eric David Petajan, 'Automatic Lipreading to Enhance Speech Recognition' (Doctoral dissertation, University of Illinois at Urbana-Champaign 1984).

⁸¹³ See generally Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier and Ashutosh Garg, 'Recent Advances in the Automatic Recognition of Audio-Visual Speech' (2003) 91(1306) Proceedings of the IEEE DOI: 10.1109/JPROC.2003.817150; Ziheng Zhou, Guoying Zhao, Xiaopeng Hong and Matti Pietikäinen, 'A Review of Recent Advances in Visual Speech Decoding' (2014) 32(9) Image and Vision Computing 590 DOI: 10/f6gqtq.

⁸¹⁴ Zhou, Zhao, Hong and Pietikäinen (n 813).

⁸¹⁵ Echoing other related fields, such as the now infamous optimism of the 1956 *Dartmouth Summer Research Project on Artificial Intelligence*, where researchers assumed AI would be a challenge with rapid progress.

ies.⁸¹⁶ In practice this is heavily augmented by context; human professionals are often referred to as speech-readers rather than lipreaders as they do not just draw on lip movements but also on surrounding factors including conversational context, facial expressions, situations and the reaction of others; meaning accurate interpretation of speech via lipreading by people who need to rely on this skill is in practice significantly more common.⁸¹⁷

As with many areas in artificial intelligence, ALRSs have recently moved from manually constructed rules and features to inference-driven machine learning.⁸¹⁸ Many recent successful models still use manually constructed features from the mouth area, for example geometric features about the mouth, descriptors of the motions observed during uttering, or features that are linked to particular articulation, like lip opening, or lip rounding.⁸¹⁹ With large enough datasets, it appears possible to automatically detect features of importance from the pixels themselves. This follows the aforementioned broad trend in ‘deep learning’, which tries to learn directly from ‘raw’ data rather than pre-processed data with humans handcrafting features thought to be predictively important.⁸²⁰ Recent techniques also try to leverage understanding of language to increase predictive power. Instead of manually segmenting sentences into words and linking those sounds to the changing shape of the mouth,⁸²¹ new approaches mimic the development of audio-only speech recognition technologies by zooming out to predicting sentences from videos, rather than extracted sounds or words from extracted mouth shapes⁸²² These techniques in the long run promise to increase accuracy by seeking to draw on a greater array of features and contexts, just

⁸¹⁶ Nicholas A Altieri, David B Pisoni and James T Townsend, ‘Some Normative Data on Lip-Reading Skills’ (2011) 130(1) *The Journal of the Acoustical Society of America* DOI: 10.1121/1.3593376.

⁸¹⁷ Helen L Bear and S Taylor, ‘Visual Speech Recognition: Aligning Terminologies for Better Understanding’ in *Proceedings of British Machine Vision Conference. 28th British Machine Vision Conference. London, UK, 4-7 September 2017, London, September 4-7 2017* (BMVA Press 2017).

⁸¹⁸ There is also some recent work with some potential use-case similarities, not utilising machine learning, around reconstruction of audio from high speed video using visible vibrations, but I do not examine these fledgling technologies here. See Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J Mysore, Frédo Durand and William T Freeman, ‘The Visual Microphone: Passive Recovery of Sound from Video’ (2014) 33(4) *ACM Trans. Graph.* 79:1 DOI: 10/gddhjm.

⁸¹⁹ The visual speech decoding research community remain considerably more divided on which features are important compared to the audio speech decoding community. See Zhou, Zhao, Hong and Pietikäinen (n 813).

⁸²⁰ See Li Deng and others, ‘Recent Advances in Deep Learning for Speech Research at Microsoft’ in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013) DOI: 10/gcsgbd.

⁸²¹ There is a second approach, not covered here explicitly but worth highlighting, where the target variable is not the word or sentence but an estimation of the audio that would have accompanied the video. See Ariel Ephrat, Tavi Halperin and Shmuel Peleg, ‘Improved Speech Reconstruction from Silent Video’ in *ICCV 2017 Workshop on Computer Vision for Audio-Visual Media* (2017) (<https://arxiv.org/abs/1708.01204>); Thomas Le Cornu and Ben Milner, ‘Generating Intelligible Audio Speech from Visual Speech’ [2017] (1751) *IEEE/ACM Transactions on Audio, Speech and Language Processing*.

⁸²² Yannis M Assael, Brendan Shillingford, Shimon Whiteson and Nando de Freitas, ‘LipNet: End-to-End Sentence Level Lipreading’ (*arXiv preprint*, 2016) (<https://arxiv.org/abs/1611.01599>).

as human 'speech-readers' do.⁸²³

3.3.3. Applications and Concerns

ALRS researchers almost exclusively cite their motivations as future user interfaces or accessibility technologies,⁸²⁴ such as automatic captioning for the deaf; silent dictation or silent speech interfaces (SSIs) for text entry or continuing a voice-synthesised phone-call in a noisy or quiet place where making a sound is either difficult or frowned-upon;⁸²⁵ as an authentication method;⁸²⁶ or simply technology to augment the performance of existing audio speech recognition technologies,⁸²⁷ Lipreading in its manual form is also already sometimes used for law enforcement and criminal evidence purposes and there will naturally be an interest in the potential for adding automated tools in this area to the barrage of CCTV and, increasingly, face recognition tools⁸²⁸ already in use in various global police forces, with associated legal controversy.⁸²⁹

While accessibility applications seem relatively uncontroversial, this is unlikely to be the case for other domains, such as law enforcement or commercial profiling. While some law enforcement use may be welcomed within limits of proportionality by most democratic populations, any technology used to add to the panoply of mass surveillance tools will be questioned as to both its legality and its morality by privacy advocates and civil liberty groups, just as CCTV, face recognition and other advanced biometric analyses such as gait recognition have already been.⁸³⁰ State surveillance law,

⁸²³ Bear and Taylor (n 817). Note that these broad trends, require considerably more labelled (ie karaoke-style subtitled) data than previous methods have, and usable datasets are scarce, and researchers report informally that access to these are restricted by perceptions of issues of consent and copyright.

⁸²⁴ See eg Assael, Shillingford, Whiteson and Freitas (n 822); John G Posa, 'Smart Phone with Self-Training, Lip-Reading and Eye-Tracking Capabilities' (*US Patent no 20130332160*) (<https://patents.google.com/patent/US20130332160>); James Vincent, 'Can Deep Learning Help Solve Lip Reading?' (*The Verge*, 7th November 2017) (<https://www.theverge.com/2016/11/7/13551210/ai-deep-learning-lip-reading-accuracy-oxford>); Joao Freitas, Ant Teixeira, Miguel Sales Dias and Samuel Silva, *Introduction to Silent Speech Interfaces* (Springer 2017) DOI: 10/gftgix.

⁸²⁵ Freitas, Teixeira, Dias and Silva (n 824).

⁸²⁶ Jiayao Tan, Xiaoliang Wang, Cam-Tu Nguyen and Yu Shi, 'SilentKey: A New Authentication Framework Through Ultrasonic-Based Lip Reading' (2018) 2(1) *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 36:1 DOI: 10/gdz6qh.

⁸²⁷ Assael, Shillingford, Whiteson and Freitas (n 822); Posa (n 824); Vincent (n 824)

⁸²⁸ See eg 'Met police to use facial recognition software at Notting Hill carnival', *The Guardian*, 5 August 2017; and see critical comments by Biometrics Commissioner in his third report, Paul Wiles, *Annual Report 2016: Commissioner for the Retention and use of Biometric Material* (Office of the Biometrics Commissioner 2017).

⁸²⁹ See eg Natasha Bernal, 'Britain's data commissioner launches investigation into UK use of facial recognition' (*The Telegraph*, 3rd December 2018) (<https://perma.cc/AS6Q-4UPX>).

⁸³⁰ In the example above, critics such as Liberty reported that 'the use of real-time biometric tracking has no basis in law and that the plan to deploy it during the carnival is institutionally racist, as it targets Britain's main annual African-Caribbean celebration'. The UK Biometrics Commissioner noted in his third annual report (Wiles (n 828)) that 'The use of facial images by the police has gone far beyond

however, is not within scope of this thesis. More relevant and least ethically appealing is the notion, so far largely overlooked by research and the media though occasionally discussed in the press,⁸³¹ that ALRSs may be extensively utilised in future for commercial purposes such as profiling and marketing. ALRSs allow analysis of high resolution videos containing many individuals, for example in public spaces or crowds. Such analysis would in theory be able to capture streams of conversations of tens, hundreds or thousands of individuals at once, insofar as enough signal, particularly lips, are visible with sufficient clarity.⁸³² Systems that can detect streams or pieces of conversations not only from individual but from *crowds* will be extraordinarily enticing for businesses engaged in retail, marketing, crowd management and security, among others, particularly when coupled downstream with text or topic mining technologies. Inevitable inaccuracies in the technique may not be detected and will likely not deter such explorations but may result in harms of various kinds to data subjects, and the inherent invisibility of post factum analysis of CCTV to which the public are already habituated means public outcry well may come too late.

Given this potential for controversy, it is worth considering how ALRSs play with the existing legislative framework: in particular, in relation to European data protection law and the lines and distinctions it has historically drawn.

This speculation about the trajectory of lipreading systems can be grounded in histories of other technologies that transform images of individuals into potentially sensitive data. Facial recognition technologies have clear parallels in that they transform datasets previously only searchable by name for validation (eg photos on driving licenses)—an easily defensible and proportionate task—into datasets that can be queried *in reverse*, with images: something significantly more controversial. A facial recognition system effectively turns video data into pseudonymous data—the trained recognition model acting as the ‘additional information’⁸³³ to link the database with the video.

Yet ALRSs are among a set of technologies designed to annotate visual data of individuals with data *other* than their name or identity (although, as discussed later, it further enables re-identification⁸³⁴). A partial comparison exists in the field of affective computing⁸³⁵, which focuses in part on automated emotion recognition such as the

using them for custody purposes’ and criticised the 19 million facial images held on the Police National Database. See also the announcement that the ICO is launching an official investigative programme into facial recognition Bernal (n 829).

⁸³¹ See eg Swindon (n 811).

⁸³² Exactly what this clarity needs to be remains a subject of academic inquiry, with little evidence that is conclusive enough to support an assessment of proportionality at this time.

⁸³³ GDPR, art 4(5).

⁸³⁴ See below at section 3.3.5.3, p. 209.

⁸³⁵ Affective computing is defined as ‘computing that relates to, arises from, or deliberately influences emotions.’ See Rosalind W Picard, *Affective Computing* (The MIT Press 1997) 249.

3. Data Protection's Lines, Blurred by Machine Learning

detection of happiness, anger, or engagement from video or wearables. Like lipreading is today, affective computing was initially presented or framed in a highly user-centric way: to make devices more subtly responsive to human needs.⁸³⁶ This too sounds uncontroversial and even desirable. Many in academia, for example, would greatly appreciate a printer that adapts to the inevitable and often regular frustration of its user (particularly when printing a document of this length).

As users of printers may have noticed, and to their detriment, affective computing saw limited application in those areas. Instead emotion detection is increasingly deployed to infer individual and collective reactions to products, places or stimuli, often covertly and without consent.⁸³⁷ Only when the automatic ordering system crashed at *Peppe's Pizza* in Oslo to reveal the command-line interface behind was it discovered (and widely reported) that a camera above the screen was passing video data to software from an American firm, *Kairos AR Inc.* who inferred and recorded the demographic, attention and affect of everyone interacting with the system.⁸³⁸ Tech companies large and small now offer APIs for individuals and small groups,⁸³⁹ while affective computing research has recently moved towards understanding characteristics of crowds,⁸⁴⁰ such as violent behaviour,⁸⁴¹ political persuasion,⁸⁴² theatre audience engagement,⁸⁴³ campus mood,⁸⁴⁴ or the 'group affect' of shared photos.⁸⁴⁵

⁸³⁶ A major founder of the field, MIT academic Rosalind Picard, notes that 'emotion has a critical role in cognition and in human-computer interaction'. Computers 'do not need affective abilities for the fanciful goal of becoming humanoids; they need them for a meeker and more practical goal: to function with intelligence and sensitivity'. Picard (n 835) 247.

⁸³⁷ For early commentary on this potential, see Joseph Bullington, "'Affective" Computing and Emotion Recognition Systems: The Future of Biometric Surveillance?' in *InfoSecCS'05* (ACM 2005).

⁸³⁸ Lisa Vaas (*Sophos Naked Security*) (<https://nakedsecurity.sophos.com/2017/05/11/would-you-like-a-side-of-facial-recognition-with-your-pizza/>).

⁸³⁹ These range from a university spin-out, *Affectiva*, to APIs from tech behemoths. See eg Microsoft, 'Azure Cognitive Services Emotion API Documentation' (*Microsoft Azure*, 27th June 2017) (<https://perma.cc/BC6Z-FY78>); Google, 'Detecting Faces: Google Cloud Vision API Documentation' (*Google Cloud Platform*, 15th August 2017) (<https://perma.cc/M82U-TAVH>).

⁸⁴⁰ Jing Shao, Kai Kang, Chen Change Loy and Xiaogang Wang, 'Deeply Learned Attributes for Crowded Scene Understanding' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)* (IEEE 2015) DOI: 10/gfrfc7.

⁸⁴¹ Tal Hassner, Yossi Itcher and Orit Kliper-Gross, 'Violent Flows: Real-Time Detection of Violent Crowd Behavior' in *2012 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (IEEE 2012) DOI: 10/gfrfc8; Mark Marsden, 'ResnetCrowd: A Residual Deep Learning Architecture for Crowd Counting, Violent Behaviour Detection and Crowd Density Level Classification' (*arXiv preprint*, 2017) (<http://arxiv.org/abs/1705.10698>).

⁸⁴² Daniel McDuff, Rana El Kaliouby, Evan Kodra and Rosalind Picard, 'Measuring Voter's Candidate Preference Based on Affective Responses to Election Debates' in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (2013) DOI: 10/gfrfc9.

⁸⁴³ Lida Theodorou, Patrick GT Healey and Fabrizio Smeraldi, 'Exploring Audience Behaviour During Contemporary Dance Performances' in *Proceedings of the 3rd International Symposium on Movement and Computing* (ACM 2016).

⁸⁴⁴ Javier Hernandez, Mohammed Hoque, Will Drevo and Rosalind Picard, 'Mood Meter: Counting Smiles in the Wild' in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp'12)* (ACM 2012) DOI: 10/gfrfdb.

⁸⁴⁵ Abhinav Dhali; Jyoti Joshi; Karan Sikka; Roland Goecke; Nicu Sebe, 'The More the Merrier: Analysing

Measuring affect is considered useful by marketers, but is ultimately a very blunt instrument compared to the potential that ALRSs offer. A model output communicated as scale or limited typology of emotions⁸⁴⁶ shines little light on the actionable reasons surrounding why a response is what it is. In advertising, some marketers are turning to ‘neuromarketing’, using neuroimaging techniques to ‘provide marketers with information that is not obtainable through conventional marketing methods’⁸⁴⁷—such as concepts they associate with something, rather than a level of satisfaction or engagement. Neuromarketing is still restricted to laboratory settings with most signals only readable through fMRI, which make it both expensive and incapable of providing information from naturalistic settings. Yet lipreading technologies at scale could promise richer ‘big data’ on crowds, suitable for both quantitative and qualitative analysis. In a high-resolution video of a shopping mall, individuals whose gaze wandered to a billboard or storefront could have sentences that followed captured and analysed on aggregate with individuals that did the same. Similarly, individuals in a clothes shop could have the product they are holding recognised, and conversations about that item recorded and processed for market research.

Indeed, similar technologies are already being used on high resolution images of crowds, such as in football stadiums. A range of relevant techniques have been published on or patented in recent years.⁸⁴⁸ South African firm Fancam (Pty) Ltd takes photos of stadiums at a 20 gigapixel resolution—equivalent to around 1,700 photos from the latest iPhone stitched together. They put these online in collaboration with the teams, but also advertise analytics services based on their image capturing technology.⁸⁴⁹ They claim⁸⁵⁰ that with their imaging and facial recognition techniques they can infer ‘actionable insights’ including:

- How many games do season ticket holders actually attend?
- How many millennials attend events and when?

the Affect of a Group of People in Images’ in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (IEEE 2015) DOI: 10/gfrfdf.

⁸⁴⁶ In relation to typologies for affect detection from facial images, see Jeffrey F Cohn and Fernando De La Torre, ‘Automated Face Analysis for Affective Computing’ in Rafael Calvo, Sidney D’Mello, Jonathan Gratch and Arvid Kappas (eds), *The Oxford Handbook of Affective Computing* (Oxford University Press 2015).

⁸⁴⁷ Dan Ariely and Gregory S Berns, ‘Neuromarketing: The Hope and Hype of Neuroimaging in Business’ [2010] (284) *Nature Reviews Neuroscience* DOI: 10/cvjds8.

⁸⁴⁸ Davide Conigliaro, Francesco Setti, Chiara Bassetti, Roberta Ferrario and Marco Cristani, ‘ATTENTO: ATTENTION Observed for Automated Spectator Crowd Analysis’ in Albert Ali Salah, Hayley Hung, Oya Aran and Hatice Gunes (eds), *Human Behavior Understanding* (Lecture Notes in Computer Science, Springer International Publishing 2013). See also <https://www.google.com/patents/US20160226610>.

⁸⁴⁹ See the webpage of Fancam as of 5 October 2017 at <http://perma.cc/EE22-SCRS>.

⁸⁵⁰ As with many recent AI business claims, these may be more indicative of marketing rather than technical progress or prowess.

3. Data Protection’s Lines, Blurred by Machine Learning

- Is our crowd aging or getting younger—and at what rate?
- What is the most optimal time to engage fans?
- Which demographic group paid the most attention to the jumbotron ad?
- Does gender distribution fluctuate, and is this fluctuation predictable or explainable?

Given the already advanced stage of imaging technology, as lipreading systems improve and become more commonplace there are few obvious insurmountable barriers to estimating a transcript for every visible individual in a crowd, simultaneously something that makes it very different from audio capture by CCTV microphones, where reported cases of their use in crowded are understandably primarily focused on detecting changes in volume rather than conversational analysis in the noise⁸⁵¹. An advantage of lipreading compared to audio speech recognition is that once a system is functioning for one speaker, having it work (to some degree) on multiple individuals in a single scene is less challenging. Although few studies have attempted it, ‘there is no reason to think that multi-person lipreading is any less viable than single-person lipreading, although the challenge of variability due to speaker identity is real’.⁸⁵² In particular, while acoustic recognition suffers from serious degradation as you add multiple speakers, visual recognition does not: as long as faces and lips can be computationally tracked and the resolution is sufficient, the number of speakers in the scene is not particularly important. If faces are able to be tracked, then some lipreading classifiers can even function on as few as two pixels-per-lip.⁸⁵³ Even if these technologies are not highly robust at scale, many commonly used methods of sentiment analysis seem equally suspect—and this has posed seemingly little barrier to their significant uptake.

3.3.4. Limitations

Lip-reading systems, like software in general, have their limits, and the reported accuracy of even the most recent breakthroughs in this space should be treated with

⁸⁵¹ Demetrius Klitou, ‘Public Space CCTV Microphones and Loudspeakers: The Ears and Mouth of “Big Brother”’ in *Privacy-Invasive Technologies and Privacy by Design* (TMC Asser Press 2014) DOI: 10/gfrfdd.

⁸⁵² Helen L Bear and Richard Harvey, ‘Phoneme-to-Viseme Mappings: The Good, the Bad, and the Ugly’ (2017) 95 *Speech Communication* 40 DOI: 10/gctgp6.

⁸⁵³ Helen L Bear, Richard Harvey, Barry-John Theobald and Yuxuan Lan, ‘Resolution Limits on Visual Speech Recognition’ in *2014 IEEE International Conference on Image Processing (ICIP)* (IEEE 2014) DOI: 10/gfrfdd. Tracking is a difficult task in lip-reading which is a significant, active area of research. Lips are non-skeletal 3D structures, while faces as a whole are complex skeleton based polygons.

scepticism. These models still struggle ‘in-the-wild’ in sub-optimal conditions of vision⁸⁵⁴ or where the participant does not always face the camera.⁸⁵⁵ Suffering as most machine learning systems do from an inability to grasp context, they struggle with the nuances of speech that hearing-impaired individuals use to reinforce their lipreading skills, such as tone or sarcasm.⁸⁵⁶

Often the most difficult challenge is to find a model which works on unseen speakers—known as ‘speaker-independent lipreading’.⁸⁵⁷ Features extracted from audio of speech are relatively speaker invariant, whilst features from video of speech are highly so.⁸⁵⁸ A phenomenon known as a ‘visual accent’ leads to a situation where ‘very similar sounds can be made by persons with very different mouth shapes’,⁸⁵⁹ with social factors a potentially mediating factor.⁸⁶⁰ Availability of high quality, naturalistic datasets is a bottleneck limiting progress on these challenges.⁸⁶¹

This section is therefore analysing a technology that is fast-moving, but one which does not *currently* have capabilities of high concern, at least without substantial inaccuracy.⁸⁶² At the time of writing there has been no published research on lipreading systems which successfully analyse crowds simultaneously. Journalistic reports that lipreading AI is now ‘better than humans’⁸⁶³ have been countered in scientific circles as largely misleading, with the media-inflated claims at best lacking in robustness and at worst blatantly disingenuous.⁸⁶⁴ But these technologies do not have to be highly ac-

⁸⁵⁴ Joon Son Chung, Andrew Senior, Oriol Vinyals and Andrew Zisserman, ‘Lip Reading Sentences in the Wild’ (*arXiv preprint*, 2016) (<http://arxiv.org/abs/1611.05358>).

⁸⁵⁵ Zhou, Zhao, Hong and Pietikäinen (n 813).

⁸⁵⁶ On the importance of such nuances, see Keith Rayner, Marcia Carlson and Lyn Frazier, ‘The Interaction of Syntax and Semantics during Sentence Processing: Eye Movements in the Analysis of Semantically Biased Sentences’ (1983) 22(358) *Journal of Verbal Learning and Verbal Behaviour*.

⁸⁵⁷ Bear and Taylor (n 817).

⁸⁵⁸ Stephen J Cox, Richard Harvey, Yuxuan Lan, Jacob Newman and Barry-John Theobald, ‘The Challenge of Multispeaker Lip-Reading’ in *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)* (2008) (<https://perma.cc/LQ9X-G6F3>).

⁸⁵⁹ Bear and Taylor (n 817).

⁸⁶⁰ On such mediation see Bernice Eisman Lott and Joel Levy, ‘The Influence of Certain Communicator Characteristics on Lip Reading Efficiency’ (1960) 51 *The Journal of Social Psychology* 419 DOI: 10.1080/00224545.1960.9922051.

⁸⁶¹ See eg Martin Cooke and Jon Barker, Stuart Cunningham and Xu Shao, ‘An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition’ (2006) 120 *The Journal of the Acoustical Society of America* DOI: 10/c7mkv9.

⁸⁶² For a summary of some of the current challenges in lipreading technologies, see Bear and Taylor (n 817).

⁸⁶³ See eg Hal Hodson, ‘Google’s DeepMind AI can lip-read TV shows better than a pro’ (*New Scientist*, 21st November 2016) (<https://perma.cc/63HM-G2D8>); BBC News, ‘AI That Lip-Reads “Better than Humans”’ (*BBC News*, 8th November 2016) (<http://www.bbc.co.uk/news/technology-3791113>).

⁸⁶⁴ See the charged debate about media hype as part of the open peer review process for Assael, Shillingford, Whiteson and Freitas (n 822) (a paper eventually rejected by the ICLR conference committee for alleged overstatement and lack of novelty of results and approach), see Various, ‘OpenReview Page for ICLR 2017 Submitted Article: “LipNet: End-to-End Sentence-Level Lipreading”’ (*OpenReview*, 1st February 2017) (<https://openreview.net/forum?id=BkjLkSqxg¬eId=BkjLkSqxg>); on the way the media exaggerated the impact of aforementioned paper, see also Vincent (n 824).

3. Data Protection's Lines, Blurred by Machine Learning

curate to be invasive. Even erroneous transcripts can betray private information if the core gist of them is correct, and can cause harm if they are not, but are taken to be. Indeed, it is worth noting that the accuracy hurdles for marketing purposes are significantly *lower* than those for user interface purposes. Users seeking to use silent dictation on their mobile devices will find it unhelpful unless it is highly reliable. Marketeers, on the other hand, are likely to be able to infer information they consider useful even with relatively high rates of transcription error.

For this reason I suggest that particular attention needs to be paid to the privacy impact of potential private sector use of automated lip reading systems. While assistive uses should be welcomed, the most immediate concern among legal and privacy academics about the use of this technology is likely be its use by government agencies for surveillance purposes, given the recent exposures, by and following Edward Snowden, of mass blanket surveillance by the NSA and other agencies. History shows however that the private use of a technology to further data collection, processing and profiling by *private* actors may ultimately be both more ubiquitous and more insidious than state surveillance use, at least in democratic societies. Indeed what Snowden has taught us is that governments can work far more effectively to surveil the masses when most the work is already done by the private sector who are then co-opted willingly or otherwise into sharing their accumulated insights, particularly around crowds. 'Surveillance capitalism'⁸⁶⁵ now works symbiotically with old fashioned state surveillance.

Additionally, as will be further discussed, the legal regime around state surveillance and investigatory powers is notably very different from data protection, and not within scope of this thesis. The main aim of this case study of lipreading systems is to examine some of the tensions within data protection law, and so limited focus is given to state uses of these technologies in favour of a focus on private use, and how existing governance regimes might apply or face challenge.

3.3.5. Regulatory challenge

Lipreading—as a human skill, rather than an automated system—has always (at least since the invention of the telescope) been recognised as inherently capable of invading privacy, with its ability to transform 'long range' sight into 'close range' overhearing of conversations. Thus, even before the computational progress discussed above, interpreter codes of practice have counselled that 'lipreading [...] videos/DVDs may constitute an invasion of privacy of the person being lipread. Using listening devices (bugs) to record conversations is unlawful without a court order: lipreading duplicates

⁸⁶⁵ Zuboff (n 17).

this concept through an alternative medium'.⁸⁶⁶

The law, at least in the UK, does not seem to have bothered itself to date with the privacy-invading quality of lipreading, probably because of its assumed position in the criminal justice system rather than a regulated activity outside of it: the only reported case relates to admissibility of lipreading testimony as criminal evidence, and refers at length to quality and consistency of the text obtained, but not as to whether any rights of privacy were invaded.⁸⁶⁷

Many technologies derived extra information by post-processing what already exists. In the case of video data, as discussed, these include approaches such as face recognition, gait recognition and affective computing. There are two distinct tensions happening. Firstly, lipreading can be seen as a continuation of the transformation of potentially non-sensitive data into sensitive data, helped along by the increasing efficacy of machine learning. But secondly, and perhaps in a newer fashion ALRSs do not just capture images, as CCTV does, or ascertain identities, as facial recognition may do, or even guess at emotional states, as affective computing can, but covertly capture *entire portions of conversations*. This shakes up data protection significantly as a prime regulatory framework supposed to govern the private sector use of these technologies in the EU, engaging a new set of issues arguably more comparable to covert audiotaping, 'wiretapping' or interception of communications, than mere video surveillance, but *without* the installation of new infrastructure such as microphones, which must operate at much closer proximity than cameras and are foiled by noisy crowds.⁸⁶⁸

The first question to consider when looking at whether ALRSs are lawful under European data protection law is whether lipreading systems process personal data. As mentioned several times in this thesis, 'personal data' is defined as 'any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly'.⁸⁶⁹ We need first to ask whether lipread conversations fall under 'any information' 'relating to' a data sub-

⁸⁶⁶ See Association of Teachers of Lipreading to Adults, 'Lipreading Interpreting Policy' (ATLA) (<https://atlalipreading.org.uk/about-us/lipreading-interpreting-policy/>) accessed 29th December 2017 ('This Code of Practice covers ATLA's position on handling requests from the media etc. for lipreaders to view and lipread events/conversations/CCTV video/DVD recordings etc. It is ATLA's policy that these requests should be refused.')

⁸⁶⁷ See *R v Luttrell (Gerrard Francis)* [2004] EWCA Crim 1344. The court held that lipreading evidence could in principle be admissible albeit with a special warning as to possible inaccuracy. See further S Jackson and G Forlin, 'Read My Lips' (2004) 154 *New Law Journal* 1146.

⁸⁶⁸ Interestingly, a recent joint European DPA project to identify the privacy risks around advanced video analytics (International Working Group on Data Protection in Telecommunications (n 810)) divided digital techniques into tools which merely detected the presence of an individual (eg shop-floor heat maps); those that classified users (eg store loyalty card schemes); and those that identified users (eg facial recognition). But ALRSs, which might capture full conversations, go beyond any of these and go into territory previously only been possible for telcos or governments with powers to order disclosure from such.

⁸⁶⁹ GDPR, art 4(1).

3. Data Protection's Lines, Blurred by Machine Learning

ject. A non-expert might intuit at first that this should merely describe characteristics such as age, sex, race, height, address, IP address etc—meta-attributes, and somehow different from full text spoken by a data subject. There is however a long history of data protection subject access requests (SARs) being used to obtain text in full relating to a data subject (eg of employment assessments⁸⁷⁰), and the recent CJEU case of *Nowak*⁸⁷¹ confirms without difficulty that text should be regarded *prima facie* as personal data. As noted there, the 'any information' part of the personal data definition 'is not restricted to information that is sensitive or private, but potentially encompasses all kinds of information, not only objective but also subjective, in the form of opinions and assessments, provided that it 'relates' to the data subject'.⁸⁷² The test for the latter, asserted by the CJEU in that case, and following the A29WP,⁸⁷³ is very wide—'it is satisfied where the information, by reason of its *content*, purpose or effect, is linked to a particular person' [emphasis added].

In *Nowak*, the data subject was seeking subject access rights to exam scripts written by himself for professional Irish accountancy exams as well as the marks given and comments written by the examiner. The CJEU was happy to hold that these scripts, though factual rather than imaginative, original or emotional, repetitive across candidates, and sometimes mathematical, 'related' to the candidate for a number of reasons including that they 'reflected [...] his intellect, thought processes, and judgment'.⁸⁷⁴ It is hard to imagine any conversation so superficial it would not similarly at least convey something about a person's tastes, habits, style of speech, native tongue or many other qualities 'relating to' that person.

More interestingly though, the *Nowak* judgment makes two points clear. Firstly, in an acrobatic feat of logic, it is important to categorise data as personal where *not* to do so might deny a data subject the right to rectify personal data which was inaccur-

⁸⁷⁰ It is interesting that one of the most notorious cases in UK data protection law, *Durant v Financial Services Authority* [2003] EWCA Civ 1746, which asserted a highly constricted definition of personal data which was restrained to data which had a 'biographical focus' on the data subject, seems to have resulted from an underlying feeling that SARs should not extend to something resembling a right to retrieve full texts. *Durant* was effectively overruled by the subsequent case of *Edem v Information Commissioner* [2014] EWCA Civ 92, even before the scope for the UK to interpret personal data differently was effectively removed by the GDPR. Still, concerns around the use of SARs in the context of 'full text' have lived on in European jurisprudence, such as in *YS and others* (n 769), where the court made it clear that data protection was not a law designed to provide access to documents, but to the information therein. See further Lilian Edwards, 'Data protection: Enter the General Data Protection Regulation' in Lilian Edwards (ed), *Law, Policy and the Internet* (Hart 2018) 85.

⁸⁷¹ *Nowak* (n 414).

⁸⁷² *ibid* ¶ 34.

⁸⁷³ Article 29 Data Protection Working Party, *Opinion 4/2007 on the concept of personal data (WP 136)* (2007). Note interestingly that the successor body to the A29WP, the EDPB have not officially adopted this as a guideline they endorse (but nor have they refuted it).

⁸⁷⁴ Interestingly they additionally held that the markers' comments also 'related' to the candidate and were thus his personal data, even though they might also be the shared personal data of the examiner.

ate or incomplete.⁸⁷⁵ This might be an important consideration for texts derived from ALRSs which are inherently likely (as discussed above), just as with a human lipreader, to contain a degree of error. The opposite claim that inaccurate text would *not* be seen as ‘relating to’ a data subject is thoroughly rejected in *Nowak*, though the court did assert firmly if without much substantive explanation that a right to correct errors would mean the right to say a paper had been falsely attributed to the wrong candidate—not to correct erroneous answers.⁸⁷⁶

Secondly, recent cases contribute to the perennial debate on ‘identifiability’ already discussed in this thesis.⁸⁷⁷ Conversations captured by video and transcribed through lipreading may arguably not *always* identify a particular data subject. Yet it seems highly likely that either generated text may contain identifying names or contextual references, or that other data collected contemporaneously (eg video images of speakers to which facial recognition algorithms might be applied) may also lead to identifiability in the hands of the data controller. As discussed below,⁸⁷⁸ this may make lipread text effectively impossible to anonymise.

However if video is generated by one party, another may apply lipreading algorithms to create speech without necessarily holding all the contextual data necessary to identify the relevant data subject who spoke the text. *Nowak*, following *Breyer*,⁸⁷⁹ and anticipating the GDPR,⁸⁸⁰ makes it very clear that text will still be regarded as personal data identifying a data subject even if the elements required to identify are held by more than one data controller. In *Nowak*, the court noted that the exam markers might not have known whose text was being marked because it was delivered to them anonymously, but it would be extremely easy for the exam organiser to later identify it using the assigned examination number. Furthermore from *Breyer* itself (which dealt with whether an IP address of a browser captured by a website is personal data) we now know for sure that data is personal in the hands of a data collector if it is reasonably likely for some data controller to have a chance of identifying it. As recital 26 of the GDPR puts it: ‘[t]o determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly’. In *Breyer*, an IP address, it was agreed, does not directly identify a person taken alone; but that piece of information will nevertheless be personal data in the hands of any party that might lawfully obtain sufficient additional data to link the information

⁸⁷⁵ *Nowak* (n 414) ¶ 53, see also GDPR, art 5(d).

⁸⁷⁶ *Nowak* (n 414) ¶ 52.

⁸⁷⁷ See section 3.1, p. 139.

⁸⁷⁸ See section 3.3.5.3, p. 209.

⁸⁷⁹ *Breyer* (n 764).

⁸⁸⁰ GDPR, recital 26.

3. Data Protection's Lines, Blurred by Machine Learning

to a person's real world identity (in that case the German government). As with IP addresses, the wisest approach is thus for any data controller to assume that lipread text produced from video is likely reasonably to be identifiable and hence personal.

An issue muddling this analysis may be that 'costs of and the amount of time required for identification' should be taken account of when deciding if a data subject is identifiable.⁸⁸¹ Yet given the leaps forward, rising accessibility and plummeting cost of reliable face recognition techniques in recent years for most data controllers not just police and intelligence services, it seems hard to assert in a robust and future-resistant manner that lipread text from video will be notably harder to reidentify than IP addresses now are.

A further question then follows: when will lip-read text generated by machine learning constitute a special category data of data (or 'sensitive personal data' as it has been known in the UK)? To recap, 'special categories'⁸⁸² of personal data in the GDPR exhaustively include racial or ethnic origins, sexuality and sexual orientation, health, political opinions, religious or philosophical beliefs, trade union membership, and genetic and biometric data processed for the purpose of uniquely identifying a natural person. It is highly likely some of these attributes will appear in everyday conversations captured on video and 're-created' as data via lipreading. Special categories of data have extra grounds required for their lawful processing⁸⁸³ and in particular, private data controllers are restricted from using the non-consensual 'legitimate interests of the data controller'⁸⁸⁴ ground for lawful processing⁸⁸⁵, or the alternate ground of declaring data 'necessary for the performance of a contract'⁸⁸⁶ and the consent ground is restricted to explicit consent only⁸⁸⁷.

If we accept then that lipread text created from captured video is best regarded as likely to be personal data, and may often even be special category data, what are the consequences? Given the limited number of other options for private controllers to process special category data, lawful processing of data to produce lipread text may become problematic for private bodies. Legality may also depend on the purpose of

⁸⁸¹ GDPR, recital 26.

⁸⁸² GDPR, art 9(1).

⁸⁸³ GDPR, art 9(2).

⁸⁸⁴ Note that GDPR, art 6(1) clarifies that *public* bodies cannot claim legitimate interest in the performance of their mandated tasks as it is 'for the legislator to provide by law for the legal basis for public authorities to process personal data' GDPR, recital 47. What is left unclear is whether private bodies can claim they are processing SPD as a matter of necessity for 'substantial public interest' GDPR, art 9(2)(g). This might provide a means whereby SPD could be processed to enable special access/deaf communities eg by subtitling cooperatives, although in the case of lipreading this is hard to imagine as a worked example, given that subtitling is usually used in the presence, rather than the absence, of audio data with a legal ground for its processing.

⁸⁸⁵ GDPR, art 6(f).

⁸⁸⁶ GDPR, art 6(e).

⁸⁸⁷ GDPR, art 9(2)(a).

processing. as the GDPR contains exemptions notably for processing related to the prevention and detection of crime.

3.3.5.1. Crime detection uses

For example, if the generated text is, as is sometimes the case already with manual systems⁸⁸⁸, used by the police to prevent, detect or prosecute crime, then their processing is exempt from the GDPR⁸⁸⁹; however regulation will then pass to the Law Enforcement Directive,⁸⁹⁰ as implemented by relevant member states, which regulates the acts of competent authorities in policing activities which involve processing of personal data. In UK law, the Directive was incorporated into Data Protection Act 2018⁸⁹¹. Under that instrument, processing of personal data by ‘competent authorities’ will be lawful where there is either consent from the data subject or it is ‘necessary for the performance of a task carried out for that purpose by a competent authority’.⁸⁹² Where special category data is processed, there must either be consent⁸⁹³ or the processing must be ‘strictly necessary for the law enforcement purpose’ *and* a condition in Sched 8 of the Act must be met.⁸⁹⁴ These conditions expand somewhat on those outlined in the Law Enforcement Directive⁸⁹⁵ and include not just judicial and statutory purposes and the administration of justice but also (*inter alia*) where necessary to detect and prevent fraud. In both cases, the competent authority must also have a policy document in place.

An interesting point however is that in the UK, unlike in some other EU member states, it has historically been accepted that private as well as public bodies can take advantage of the crime exemption, in part fuelling the exponential rise in high street, mall and residential areas CCTV surveillance in the UK and its reputation as surveillance capital of the EU⁸⁹⁶. We can therefore imagine two scenarios involving lipreading systems in particular:

In scenario A (drawn from real life but adapted), the Scottish Professional Football League (SPFL) decides to not only utilise CCTV at its games to inhibit disorder and have

⁸⁸⁸ See *Lutrell* (n 867) discussed above

⁸⁸⁹ GDPR, art 2(1)(d).

⁸⁹⁰ Law Enforcement Directive.

⁸⁹¹ DPA 2018, part 3.

⁸⁹² *ibid* s 35(2).

⁸⁹³ *ibid* s 35(4).

⁸⁹⁴ *ibid* s 35(5).

⁸⁹⁵ Law Enforcement Directive, art 10.

⁸⁹⁶ Note that domestic CCTV may sometimes also benefit from the ‘household exemption in GDPR, art 2(2)(c)—however this will not extend to CCTV where cameras located on domestic premises capture video from public areas. See Case C-212/13 *František Ryneš v Úřad pro ochranu osobních údajů* ECLI:EU:C:2014:2428.

3. Data Protection's Lines, Blurred by Machine Learning

evidence to pass to the police if it breaks out, but also decides to pass video through an ALRS to determine if there are any breaches of the Scots law which criminalised menacing communications and the making of racist remarks, including the chanting of sectarian songs, at sporting events.⁸⁹⁷

If the SPFL is a 'competent authority' then it might claim the prevention of crime exemption, which would mean it was exempted from the need to show a lawful ground of processing under the GDPR and would only need to comply with the Law Enforcement Directive as implemented in part 3 of the DPA 2018. A 'competent authority' is defined as one of a limited list of public bodies enumerated in Sched 7 or 'any other person if and to the extent that the person has statutory functions for any of the law enforcement purposes'.⁸⁹⁸ Whether this fits the SPFL, is not yet clear: while the SPFL is actively engaged with Police Scotland in preventing crime and disorder at matches and in general within the football ecosphere,⁸⁹⁹ it itself is not a law enforcement body but a sports regulator. If the SPFL was to be so designated however, processing of CCTV captured at its matches would seem lawful where it was 'strictly necessary for the law enforcement purpose'—which remains a high level of test given the argument that ordinary video capture plus face recognition might have sufficed. This does not seem like a get-out-clause for the SPFL, nor does it seem very feasible for other private actors or sectors more widely.

What, however, if the video captured and processed via an ALRS was also used for other purposes, to do with marketing and increasing profits, such as to get an idea of fan satisfaction with the game, the catering, the stadium, and access and egress arrangements?⁹⁰⁰ Those non-law enforcement related activities are considered next.

3.3.5.2. Non-crime detection uses

In scenario B, a shopping mall uses a variety of techniques including WiFi hotspot data logging, phone signals, smart sensors (eg on doors, floors), and smart CCTV cameras to measure footfall in the mall and provides tenants with data as to how customers

⁸⁹⁷ This has been adapted from the discussion of face recognition from CCTV and its potential use by the SPFL (hampered by lack of central funding to develop the scheme) at BBC News, 'SPFL facial recognition cash blow from Scottish Government' (*BBC News*, 25th February 2015) (<http://www.bbc.co.uk/sport/football/35664117>). Note that while all these activities were potentially criminalised by the Offensive Behaviour at Football and Threatening Communications (Scotland) Act 2012, during writing this Act was repealed. Given the topic in question is not football regulation, it is retained here for illustrative purposes.

⁸⁹⁸ DPA 2018, s 30.

⁸⁹⁹ See activities discussed at <http://www.scotland.police.uk/whats-happening/featured-articles/behind-the-scenes-football-coordination-unit-for-scotland>, including the activities of FoCUS, the Football Coordination Unit for Scotland.

⁹⁰⁰ The *Fancam* example above (p. 199) illustrates a company claiming to provide these services to sports stadiums today.

moved around, how often they paused in front of and entered various shops, at what times of day, their demographic groups etc.⁹⁰¹ Notices in the mall advise that CCTV cameras are in operation but that data is only used for law enforcement purposes, and the full privacy policy posted on the website asserts that data is also used to provide aggregate marketing intelligence but all personal data processed is anonymised.

In this scenario, data is clearly being used for commercial purposes either instead of or as well as law enforcement purposes. It is likely the mall operator would offer this intelligence to its tenants for a premium. As noted in the earlier section, this is very likely if not already in progress. In Glasgow, for example, considerable press resulted in 2017 from a security firm's suggestion that with lipreading CCTV data, businesses will '[...] analyse their customers' reactions to a particular aspect of their store or service, and get qualitative as well as quantitative data. By capturing the comments of customers this way, they get an insight into unedited and genuine information that couldn't be captured any other way'.⁹⁰² Shopping centres, it is claimed, can obtain more genuine and more usefully targeted data from lipreading than they can from shopper surveys, focus groups or other existing means of sampling opinion.

As the likely collection of special category cannot be securely ruled out, and no practical means exists to collect explicit consent from those surveilled in public spaces like malls to lift the special category processing ban, the strategy for lawful processing here can *only be* to claim that data is anonymised. Is this strategy realistic? It is already commonplace to claim that this is done,⁹⁰³ but is facing serious challenge in an era of big data, profiling and re-identification. How does it stand up if we posit that the mall also begins to generate lip-read data from the high quality video it captures?

3.3.5.3. The (im)possibility of anonymisation

Let us compare anonymisation of spoken transcripts generated manually from audiotaped speech. This is currently commonly undertaken by qualitative researchers, who seek to reduce the identifiability of individuals they have interviewed, particularly if those interviews concerned sensitive discussions, such as medical histories.⁹⁰⁴ In order to preserve data as identifiable and useful, the standard practice is to replace names with codes (eg <F01> for a female name) or equivalent value (eg a different

⁹⁰¹ See generally Mavroudis and Veale (n 611).

⁹⁰² Swindon (n 811). Interestingly, in relation to Scenario A, the article also notes that although we [...] generally associate this type of functionality with interpreting what contentious things were said by players at football matches, it has far wider-reaching implications, especially for business.'

⁹⁰³ The anonymisation card was also 'played' in this manner by TfL, see section 3.1.3.

⁹⁰⁴ Benjamin Saunders, Jenny Kitzinger and Celia Kitzinger, 'Anonymising Interview Data: Challenges and Compromise in Practice' (2015) 15(5) *Qualitative Research* 616 DOI: 10/f7sk3r.

3. Data Protection's Lines, Blurred by Machine Learning

name associated with age group, social class or region).⁹⁰⁵ These techniques are considered acceptable but only in interview situations where participants are aware they are being recorded, have given ethical consent,⁹⁰⁶ and can decide what to share and what to omit. In this situation, identifiability and explicit consent are both possible and compatible, and anonymisation takes place after, not during, collection. (Note also that in the GDPR such substitution would be regarded as pseudonymisation not anonymisation, which explicitly means the data remains personal⁹⁰⁷.)

By contrast, in the shopping mall scenario, video is captured at a distance, from crowds not individuals, possibly in some time period separated from later analysis, possibly by a different party to the one who captured the original video, and without opportunity to take explicit consent releases. In practical terms, only warning by notice is possible and it is dubious if this 'take it or leave the mall' approach can possibly constitute consent to processing of ordinary personal data—let alone explicit consent to processing of sensitive personal data—under the GDPR, where consent is to be given by 'a clear affirmative act'⁹⁰⁸ and silence is not an acceptable surrogate, and where consent is 'presumed not to be freely given [...] if the performance of a contract, including the provision of a service, is dependent on the consent despite such consent not being necessary for such performance'.⁹⁰⁹

Thus the only option for private data controllers processing data acquired via ALRSs for marketing purposes, then, is to claim (1) that their technology either renders data fully anonymised or at least (2) excludes generation of sensitive personal data aspects. In the former scenario, they can escape data protection law entirely; in the latter scenario, a reasonable claim of lawfulness via legitimate interests might be posited.⁹¹⁰

⁹⁰⁵ Frances Rock, 'Policy and Practice in the Anonymisation of Linguistic Data' (2001) 6(1) *International Journal of Corpus Linguistics* 1 DOI: 10/cg534s.

⁹⁰⁶ Depending on the country, participants may not need to give data protection consent, even if disclosing SPD. The UK for example allows public universities to place research under the 'public task' ground for processing, so consent is generally sought in line with good ethical practice and for compliance with other reasons, such as (in England and Wales), the common law duty of confidentiality. See Medical Research Council, *General Data Protection Regulation (GDPR): Consent in Research and Confidentiality* (UKRI 2018) 5.

⁹⁰⁷ See GDPR, art 4(5) and GDPR, recital 26. While pseudonymised data qualifies for some regulatory exemptions as a type of privacy by design, largely the normal rules of data protection continue to apply.

⁹⁰⁸ GDPR, recital 32.

⁹⁰⁹ GDPR, recital 43. Entering a mall *per se* is not assumed to involve the making of a contract. However consent might conceivably be collected retrospectively via shared terms and conditions when/if purchases are made at tenant shops.

⁹¹⁰ For an early examination of a parallel to this strategy in the context of profiling, see Wim Schreurs, Mireille Hildebrandt, Els Kindt and Michaël Vanfleteren, 'Cogitas, Ergo Sum. The Role of Data Protection Law and Non-discrimination Law in Group Profiling in the Private Sector' in Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer 2008) DOI: 10/ccpqh5.

3.3.6. Privacy by design as an escape route?

An interesting analysis can be found in so far the only UK case reported about legitimate non-police use of audio surveillance in the context of DP and privacy rights. The case concerned the recording of both passengers and drivers required in licensed taxi cabs in Southampton by the local council as licensing authority. The ICO released a binding enforcement decision requiring the practice to cease,⁹¹¹ which was contested but upheld by the Information Rights Tribunal.⁹¹² The requirement was transparently intended purely to better record and thus dissuade disorderly behaviour in cabs for the protection of both passengers and drivers. This case has both data protection elements as well as concerning European Convention on Human Rights.

While the Information Commissioner had accepted that the policy served a legitimate aim and that there was a pressing social need for some surveillance in taxis, and in some cases, audio recording had made a real difference to protecting drivers over and above CCTV eg regarding racist remarks, Southampton's policy of continuous audio-recording was simply not 'proportionate', given that it meant 'every single conversation, however private and however sensitive the subject matter, taking place during every single taxi ride in Southampton (of which there may well be a million a year) will be recorded and accessible to a public authority'. The 'marginal benefits to the legitimate social aims of increasing public safety and reducing crime in relation to taxis which were likely to result from it' were not enough to justify it. Interestingly in light of previous discussion, the Tribunal was particularly concerned that special category data would almost invariably be collected. It was 'quite satisfied that the inhabitants of (and visitors to) Southampton will from time to time discuss their own and others' sex lives, health, politics, religious beliefs and so on in taxis (notwithstanding the presence of the taxi driver) and, if necessary, we take judicial notice of that fact', noting also that sensitive matters that did not fit the restrictive categories of special category data under data protection law would also be aired, which even if not sensitive in data protection would have relevance in relation to article 8 of the European Convention on Human Rights (ECHR).

Fascinatingly though, as a final note the Tribunal speculated that what we might call a privacy by design (PbD) solution could be possible, including, for example, the use of panic buttons to activate audio recording only when desired, disabling of such tech when a contract to carry a child or vulnerable adult was in operation, etc. How far might PbD approaches, applied to ALRSs, help data controllers claim their two main exit routes above: that i) the data is anonymised, or ii) that special category data is not

⁹¹¹ See the enforcement notice at <https://perma.cc/3L82-338P>.

⁹¹² *Southampton City Council v Information Commissioner* [2013] UKFTT 20120171 (GRC).

captured?

3.3.6.1. Data minimisation

Data minimisation is the notion of only collecting and retaining data 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed'.⁹¹³ In terms of video surveillance, the word minimum will often be taken to mean the *quantity* of data being stored, or the number of variables, but might also refer to the *quality* of the video collected which may restrain the kinds of post-processing I have been discussing. Only recording what is necessary is already an obvious DPbD principle—one emphasised in both *So'ton CC*, and in the ICO's CCTV guidance⁹¹⁴—but in crowded spaces this can be practically challenging. However the main risks lie in the *qualities* of the data collected. Computer scientists often talk about the *dimensionality* of datasets—roughly speaking, how much non-redundant information is present for each observation. Often, this is thought of in terms of attributes, like columns in a dataset, but transcripts are data where attributes are somewhat hidden, and only revealed when analytic methods are applied (which transform textual data into something like the traditional 'columns are attributes' model). Data minimisation in this context should therefore consider how it reduces the number of different attributes that can be inferred from data analysis (as all text data quantitatively analysed will be processed in this way), in addition to more traditional conceptions such as reducing the length of video footage captured.

Data minimisation techniques for video data in relation to lipreading potential should be developed with some urgency, particularly to aid controllers capturing data with no intention to perform lipreading. While I am not aware of any technologies specifically designed for this purpose, several recent technological developments give clear hope that such a system is not out of reach. Blurring faces is common but this may be aesthetically undesirable, and damage the commercial value of video data. Instead, it might be possible to replace the lip movements with something appearing visually similar to a viewer, but which contains considerably less, or even no, retrievable data about what was said. Tools already exist for manipulating lips realistically into chosen, arbitrary speech,⁹¹⁵ and other areas of image recognition are develop-

⁹¹³ GDPR, art 5(c).

⁹¹⁴ Information Commissioner's Office, *In the picture: A data protection code of practice for surveillance cameras and personal information (v1.2)* (ICO 2017) (<https://ico.org.uk/media/1542/cctv-code-of-practice.pdf>).

⁹¹⁵ Supasorn Suwajanakorn, Steven M Seitz and Ira Kemelmacher-Shlizerman, 'Synthesizing Obama: Learning Lip Sync from Audio' (2017) 36(4) ACM Trans. Graph. 95:1 DOI: 10/gdgpz4; Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt and Matthias Niessner, 'Face2Face: Real-Time Face Capture and Reenactment of RGB Videos' in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

ing tools to attractively censor different areas of an image that might betray sensitive information.⁹¹⁶ This is promising, as intuitively it seems a more difficult task than manipulating lips into speech which *appears* realistic, but is not constrained by having to *be* realistic.

Secondly, controllers that *do* wish to undertake lipreading analysis on data they collect or hold must consider whether the transcripts they obtain only contain data which are ‘relevant and limited to what is necessary’.⁹¹⁷ Given that this data is textual, it is not quite as simple as removing unnecessary variables. Most non-trivial aspects of textual analysis (eg other than aspects such as character counts) require further processing, usually with external datasets. In particular, textual analysis (eg natural language processing) requires the transformation of a transcript into a form more statistically digestible, such as a matrix. Commonly, blocks of text are turned into a data construct called a term- or word-document matrix.⁹¹⁸ Each row of these matrices represents a particular word or phrase that is in a preset vocabulary (eg things that have been said in earlier training data), while each column represents a ‘document’, which in this case could be a sentence or a longer comment. These ‘documents’ are transformed into a vector describing numerically which words and phrases were present, and which were not. A common task that follows is classification: a machine learning model takes the document vector as input, and returns a prediction, such as the type of document it is.

This process has the potential to be carried out in a data minimising manner. In general, only the words in the selected vocabulary are used; other words mentioned are discarded. If there is no row in your existing matrix that represents a word, it is ignored. This is particularly important if it is chosen to use a model, such as a pre-trained classifier, as this has a predefined vocabulary and can normally only consider words it already knows (it cannot infer anything from new terminology beyond that unknown words are present). In practice this means a list of words exists that can be minimised in advance to avoid blatant privacy breaches from rareness or specificity (although, of course, people can express highly private facts using only highly generic words). For example, most proper nouns, which could potentially represent individuals or employers, could be dropped by default—except for those of interest, such as the names of brands, shops or certain landmarks. Because data can be stored in such a matrix, there is no need to retain the transcript after transformation.

⁹¹⁶ See eg Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J Crandall, Roberto Hoyle and Apu Kapadia, ‘Viewer Experience of Obscuring Scene Elements in Photos to Enhance Privacy’ in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18, ACM 2018) DOI: 10/gfkr9s.

⁹¹⁷ GDPR, art 5(c).

⁹¹⁸ Christopher D Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing* (MIT Press 1999).

3.3.6.1.1. Sensitive data minimisation Connected to data minimisation in the context of attributes 'contained' within text is the idea that controllers should, if not relevant, limit or avoid the use of special category data⁹¹⁹, as well as use techniques to limit the re-identification of individuals⁹²⁰. Both these strategies were noted above already as likely claims to legitimise processing on ALRSs earlier in the section on regulation.

The A29WP has emphasised in guidance⁹²¹ that 'profiling can create special category data by inference from other data which is not special category data in its own right but becomes so when combined with other data'. They warn that where such data is inferred, it should legally be treated identically as special category data directly provided or observed directly. But being able to anticipate the production of special category data implicitly assumes a supervised machine learning model trained explicitly with input variables linked to sensitive data. With textual data analysis, however, typically certain phrases and declarations (eg somebody claiming that they had cancer) are not encoded directly as a variable in the data, but are also not inferred opaquely through statistical correlations. Indeed, systems that process claims and logics from the structures of text are designed to answer questions using just the given text without reference to other sources of data about an individual.⁹²² Special category data in textual data is therefore less about *combining* datasets, and more about asking questions of the data you already have. If we cannot transform this data to resist questions that lead to the production of special category data, controllers could be accused of holding it in an easily accessible form, which might endanger the grounds for processing they are relying on, particularly legitimate interests or consent which is not 'explicit'.⁹²³

The simplification and redaction of data in a term-document matrix, discussed above as a form of data minimisation, may equally be helpful here. But removing the grammatical order, larger vocabulary and context which could be used for more powerful analysis as natural language processing methods develop may be highly destructive to future uses. Is it possible to edit the original transcripts, leaving them largely intact, but with clearly sensitive components removed?

Some inspiration can be drawn from computational domains which attempt to 'sanitise' sensitive documents, such as written medical records, for the purposes of research.⁹²⁴ However not only are these systems not designed for spoken transcripts,

⁹¹⁹ GDPR, recital 51.

⁹²⁰ GDPR, recital 26; GDPR, recital 28; GDPR, recital 29.

⁹²¹ Article 29 Data Protection Working Party, *ADM Guidelines* (n 2).

⁹²² Rudolf Kadlec, Martin Schmid, Ondřej Bajgar and Jan Kleindienst, 'Text Understanding with the Attention Sum Reader Network' in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics 2016) DOI: 10/cxkt.

⁹²³ As per the requirements in GDPR, art 9.

⁹²⁴ See eg Venkatesan T Chakaravarthy, Himanshu Gupta, Prasan Roy and Mukesh K Mohania, 'Efficient

they are designed for well-structured tasks, primarily medical tasks, with a delimited number of sensitive characteristics (eg granularity of disease) which are of concern. Those may be amenable to removal, but the definition of special category data in the GDPR is much wider, covering aspects such as political opinion which are hard to mitigate given they might be inferred through contextual statements⁹²⁵. This reveals a difficult grey area. While the statement of “I have been diagnosed with cancer” would probably be easily pegged as containing special category data, what about the phrase “I am on cisplatin”? Cisplatin is a common chemotherapy drug, but assuming that the speaker has been diagnosed with cancer is nevertheless an act of inference, albeit one which could be achieved with generic dictionary look up rather than requiring any further information about that specific individual. It will be hard for controllers attempting to sanitise their datasets to avoid disclosing SPD to know how far such phrases have to be removed and privacy by design approaches such as sanitisation will likely struggle to deal with sensitive disclosures which appear obvious to humans but can be easily missed by computer systems.

3.3.6.1.2. Anonymisation Connectedly, individuals can also disclose information which can be used to identify or re-identify them. While visual data is not automatically considered as biometric SPD in the GDPR, it can be after it is processed by technical means heightening the ability to re-identify⁹²⁶. An individual naming an employer, address, location, or speaking their contact details clearly makes their connected transcript at high risk of re-identification; yet automatically removing all these varied forms of data with high accuracy is a challenging task indeed⁹²⁷. The ‘style’ of speaking is also something which controllers may want to minimise or mitigate, as stylometric (or linguistic) analysis of text has been used by several security and privacy researchers to mount reidentification attacks on text. Such attacks include attempting to detect the distance between two authors’ ‘writeprints’,⁹²⁸ or by comparing all pairs of potential authors together in a “doppelgänger finder” approach⁹²⁹ to cluster

Techniques for Document Sanitization’ in *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (CIKM ’08, ACM 2008) DOI: 10/frzwmc; David Sánchez, Montserrat Batet and Alexandre Viejo, ‘Minimizing the Disclosure Risk of Semantic Correlations in Document Sanitization’ (2013) 249 *Information Sciences* 110 DOI: 10/f5b2pn; David Sánchez and Montserrat Batet, ‘C-Sanitized: A Privacy Model for Document Redaction and Sanitization’ (2016) 67(1) *Journal of the Association for Information Science and Technology* 148 DOI: 10/f77swg.

⁹²⁵ See Article 29 Data Protection Working Party, *ADM Guidelines* (n 2).

⁹²⁶ GDPR, recital 51.

⁹²⁷ David Sánchez and Montserrat Batet, ‘Toward Sensitive Document Release with Privacy Guarantees’ (2017) 59 *Engineering Applications of Artificial Intelligence* 23 DOI: 10/f9s634.

⁹²⁸ Ahmed Abbasi and Hsinchun Chen, ‘Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace’ (2008) 26(2) *ACM Trans. Inf. Syst.* 7:1 DOI: 10/dq78jg.

⁹²⁹ Sadia Afroz, Aylin Caliskan Islam, Ariel Stoleran, Rachel Greenstadt and Damon McCoy, ‘Doppelgänger finder: Taking stylometry to the underground’ in *2014 IEEE Symposium on Security and Privacy*

all spoken references by a single individual. All this seems to indicate that an individual's distinctive spoken style as transcribed is as identifying and revealing, if not more so, than their written style, and consequently that anonymisation of speech texts is an extremely difficult job.

3.3.6.2. Purpose limitation

The overarching principle of purpose limitation⁹³⁰ is that data collected for one purpose should not be re-used for another that is incompatible with the original; at least without some fresh lawful ground for processing. Yet re-use is quintessentially what ALRSs enable, given pre existing high quality video. Technically however, we can ask if it is possible to 'bake' purpose limitation into the collection and transformation of visual data so that it cannot be easily misused.

There are two main technical approaches I see as relevant here. The first consists of ensuring the analysis undertaken aggregates data in a form where it is difficult to use for more granular purposes. One way to do this is to aggregate data in a way that deindividualises it.⁹³¹ If the target of analysis is truly the *crowd* (eg collective customer opinion over an entire day) rather than the *individual*, individual level data should be discarded as soon as technically feasible in a way that maximises analytic possibilities at the crowd level, but without allowing individuals to be singled out again.⁹³²

A second, perhaps less conventional approach, consists of bundling the hardware and the analytical software together in such a form that the only visible output to the data controller is a classification or score, rather than the raw or otherwise recognisable transcripts. An example is a product from the *Fraunhofer Institute for Integrated Circuits IIS* in Germany, which has developed facial analysis software called SHORE which has been used by researchers and practitioners to analyse the behaviour of crowds in public spaces.⁹³³ SHORE claims to be able to estimate gender, age, as well as four facial expressions: happy, sad, surprised and angry, as well as the state of various facial features.⁹³⁴ As facial analysis is carried out over personal data; and potentially spe-

(SP) (2014) DOI: 10/cwnz.

⁹³⁰ GDPR, art 5(1)(b).

⁹³¹ Note that aggregation, if done poorly, can also be vulnerable to re-identification, particularly where small groups are concerned. See generally William E Winkler, 'Re-Identification Methods for Masked Microdata' in Josep Domingo-Ferrer and Vicenç Torra (eds), *Privacy in Statistical Databases* (Lecture Notes in Computer Science, Springer Berlin Heidelberg 2004) DOI: 10/d2jx8t.

⁹³² A blurred line exists between this and data minimisation as discussed above, as if data is aggregated in a way that is irreversible, it would be considered anonymous and outside of the scope of data protection law.

⁹³³ See eg Theodorou, Healey and Smeraldi (n 843).

⁹³⁴ Fraunhofer Institute for Integrated Circuits, 'Face Detection Software SHORE: Fast, Reliable and Real-time Capable' (*Fraunhofer IIS*, 2018) (<https://perma.cc/9PWR-QGK6>) accessed 3rd December 2018.

cial categories of personal data where they reveal factors such as race, or are biometric in the context of identification,⁹³⁵ they sell an entire workflow called Anonymous Video Analytics for Retail and Digital Signage (AVARD) designed to make the technology compliant with data protection law. This integrates the analytic software inside a tamper-resistant black box which includes the camera⁹³⁶, and therefore according to the Institute ‘merely transmits metadata or statistics’, for example, about a visitor’s perceived sentiment.⁹³⁷ As a result, they have sought and received third party certification that claims adherence with DP law for this product, and claim that ‘the Bavarian DP Authority for the Private Sector (BayLDA) confirms that the AVARD system doesn’t process personal data but only anonymous data’.⁹³⁸ Indeed, previous work on CCTV audio has floated similar ideas of triggering recording only upon the detection of certain words—presumably, while the microphone is constantly recording, data not meeting detected criteria are quickly discarded.⁹³⁹

Challenges remain with this direction. Firstly, it is unclear to what extent the BayLDA opinion is consistent with the law: to claim that no personal data is processed simply because the output is aggregate rather than personally identifying seems counter-intuitive. Much will depend on how far input data could be reconstituted and/or individuals re-identified. I leave detailed legal analysis of this setup for future work. Secondly, such a setup would make it difficult to train models in the future, as new training data (which could potentially be labelled) would not be retained. In this case, the machine learning systems analysing the data would have to be bought in from elsewhere, and they themselves would need adequate legal grounds for processing were they to be trained in Europe.

3.3.6.3. Objection

Signalling objection, an important GDPR data subject right,⁹⁴⁰ is difficult where biometric data is collected pervasively, and your data record is not easily or reliably associated to a name or other persistent identifier. Unilaterally seeking to object without

⁹³⁵ GDPR, art 9(1).

⁹³⁶ While it is outside the scope of this paper, we point to developments in secure hardware enclaves and secure computation that might, in theory, provide mathematical safeguards for these systems that lead them to be credibly tamper-resistant. Indeed, it could be envisaged that these systems could be subject to the certification provisions of the GDPR, art 42.

⁹³⁷ Fraunhofer Institute for Integrated Circuits, ‘Fraunhofer IIS Presents an Intelligent Sensor for Anonymous Video Analysis at the MWC in Barcelona’ (27th February 2018) (<https://perma.cc/Y3VS-59VZ>) accessed 3rd December 2018.

⁹³⁸ The report (LDA-1085.4-1368/17-I, dated 8 June 2017) was obtained by the author upon request from the Fraunhofer Institute, who have requested it not be shared.

⁹³⁹ Klitou (n 851).

⁹⁴⁰ GDPR, art 21. See further section 3.2.5.3.

3. Data Protection's Lines, Blurred by Machine Learning

having to trust the operator to obey your preferences is particularly challenging as it usually requires a visible and rather odd change in behaviour.⁹⁴¹ Methods do exist to 'fool' visual surveillance systems or analytics systems, but they currently require wearing specific, often silly-looking, clothing,⁹⁴² and are both unlikely to catch on and provide few, if any, lasting guarantees of protection—particularly as systems will then be trained to overcome the confusion they offer.

Some novel opt-out methods have been proposed by technology platforms which might have relevance in this setting, particularly those adopted to prevent circulation of images of illegal or unpleasant abuse eg 'revenge porn'. Platforms such as Facebook can easily hold copies of illegal images and match them against uploaded images, blocking the upload where a match is recorded. However platforms do not wish to hold large databases of these images for both reasons of reputation and security, and because possession may itself be illegal. Accordingly, specially adapted 'hashing' methods have been used to overcome this. Hashing any file—an image, a block of text or code—transforms it in a one-way process into a shorter mathematical representation of the file. Hashing is a form of technical accountability as described earlier in this thesis.⁹⁴³ Anyone applying a certain hash function to the same file will get the same hash. Usually, because the hash is much shorter than the original file, this is a quick method of verifying that two people are looking at the same code.⁹⁴⁴ Because it is one-way, and betrays no information about the file in question beyond the ability to match, adapted versions of these techniques (designed to be resistant to minor changes in the underlying images) are used to 'store' illegal images, retaining the ability to match them against new comparator images without the ability to view them.⁹⁴⁵

Hashing an image might be adapted as an opt out method for lipreading systems too. 'Robust hashing' techniques have been designed for facial data in particular, where it might be desirable to recognise that a face is (or is not) on a certain list without having a reconstructable list of faces.⁹⁴⁶ Other systems assume that a full version of the face

⁹⁴¹ See further Mavroudis and Veale (n 611).

⁹⁴² See eg Mahmood Sharif, Sruti Bhagavatula, Lujó Bauer and Michael K Reiter, 'Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition' in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016) DOI: 10/bsmm; Takayuki Yamada, Seiichi Gohshi and Isao Echizen, 'Use of Invisible Noise Signals to Prevent Privacy Invasion Through Face Recognition from Camera Images' in *Proceedings of the 20th ACM International Conference on Multimedia* (MM '12, ACM 2012) DOI: 10/gfksck.

⁹⁴³ See section 1.6.3, p. 71.

⁹⁴⁴ See further footnote 618.

⁹⁴⁵ See eg the *PhotoDNA* software developed for child pornography detection by Microsoft and Hany Farid.

⁹⁴⁶ Yagiz Sutcu, Husrev Taha Sencar and Nasir Memon, 'A Secure Biometric Authentication Scheme Based on Robust Hashing' in *Proceedings of the 7th Workshop on Multimedia and Security* (ACM 2005) DOI: 10/cj8sj5; Marta Gomez-Barrero, Christian Rathgeb, Javier Galbally, Julian Fierrez and Christoph Busch, 'Protected Facial Biometric Templates Based on Local Gabor Patterns and Adaptive Bloom Filters' in *Proceedings of the 2014 22nd International Conference on Pattern Recognition* (ICPR '14, IEEE Computer So-

is stored, but on a trusted server, and an intermediary (like a shopping mall) is able to undertake a computation with server that returns the answer to the question ‘is this person on this opt-out list’ without either the camera-operator learning more about the individual, or the server knowing who was being queried.⁹⁴⁷ Naturally, while cryptographers may trust these systems, it would remain to be seen whether users would feel they provide assurance, particularly when they must provide biometric data to the third party and, quite literally, take their assurances of irreversibility at face value.⁹⁴⁸

Such technologies may be interesting avenues to explore. Such a system could run faces through this recognition system as an intermediate step, and when the system returns either a positive or a negative result, use that to determine whether the patterns from that face are analysed—or, indeed, whether their lips are obfuscated completely in any stored footage.

3.3.7. Interim discussion

Lipreading technologies both blur the lines between sensitive and non-sensitive data as well as blurring the concept of what data is, could, or should be considered ‘sensitive’. Machine learning aids in re-identification, in exacerbating the sensitivity of transcripts and recordings, and the transformation of data in ways which violate data subjects’ expectations. It both challenges the idea that the ‘sensitive’ data of the GDPR captures what we societally feel is sensitive, but also highlights the blunt nature of the tools that exist, and how lipreading systems might with high likelihood transform much of the data they touch into sensitive personal data, requiring explicit consent (for most actors) in order to permit processing. PbD provides some useful tools, such as technical means to minimise, limit purposes or to object, but these are not only stressed by the high dimensional nature of the data in question, which contains plenty of implicit information, but which also beyond providing safeguards to little to resolve the underlying tensions in data protection that lipreading technologies help surface.

ciety 2014) DOI: 10/gfksen.

⁹⁴⁷ Peter Aldhous, ‘The Digital Search for Victims of Child Pornography’ (2011) 210(2807) *New Scientist* 23 DOI: 10/b3brn; Ahmad-Reza Sadeghi, Thomas Schneider and Immo Wehrenberg, ‘Efficient Privacy-Preserving Face Recognition’ in Donghoon Lee and Seokhie Hong (eds), *Information, Security and Cryptology – ICISC 2009* (Lecture Notes in Computer Science, Springer Berlin Heidelberg 2010).

⁹⁴⁸ Opponents of the Facebook hashing scheme have pointed out that users could prepare the hash of their own image at home using an app and merely upload it to Facebook. One wonders if this could work as an authenticated opt-out to CCTV capture when combined with some kind of digital signature.

3.4. Summary remarks

In this chapter, I have considered how machine learning and data protection law mix *beyond* the question of which rights apply and how. I highlighted several core assumptions and tensions which had not been discussed previously: that controllers could shift risk onto data subjects by deleting explicit data subject identifiers, thus denying individuals their rights; that machine learning models might themselves be personal data; and that some transformations of data, such as those undertaken by lipreading systems, shake data protection's notions of what is 'sensitive' and what is not.

Data protection is, for the most part, a robust regime which deals with a range of technologies in a flexible manner. Yet some of the issues and characteristics of machine learning technology and practice do stress the framework. Some of these stressed areas should be priority areas for debate discussion if and when the GDPR is revised or (in part) copied to other jurisdictions. They do not come with clear solutions but, as has been laid out, are beset with trade-offs. This chapter has attempted to provide a guide to these areas, and the work presented above motivated in such a way as to stimulate policy debate in the years to come.

Part III.

Machine Learning on the Ground

4. Coping with Value(s) in Public Sector Machine Learning

Public bodies and agencies have increasingly sought to use new forms of data analysis to provide ‘better’ public services. These reforms have included digital service transformations such as ‘e-government 2.0’ and the creation of ‘integrated data infrastructures’ (linked administrative datasets),⁹⁴⁹ generally aimed at ‘improving the experience of the citizen’, ‘making government more efficient’ and ‘boosting business and the wider economy’.⁹⁵⁰

It is far from a new observation that administrative data—data collected by or for public bodies for registration, transaction and record keeping—might be mined for better understanding of societal patterns, trends and policy impacts, or sanitised and released to fuel innovative products and services. A plethora of government reviews and initiatives have, especially over the last decade, led to the establishment of centres, networks and infrastructures (such as the UK’s Administrative Data Research Network) to better understand societal phenomena using these data sources.⁹⁵¹ Yet more recently, there has been a push to use administrative data to build models with the purpose of helping make day-to-day operational decisions in the management and delivery of public services, rather than providing general evidence to improve strategy or government-citizen interaction.

These new operational models are designed to serve as decision support or even to trigger automatic action, and area often built with machine learning components that range from simple to sophisticated.

Information technology is supposed to be a ‘central force’ to transformations in public management,⁹⁵² although despite decades of promise of transformation, these

⁹⁴⁹ Statistics New Zealand, ‘Integrated Data Infrastructure’ (*Government of New Zealand*, 2016) (<https://perma.cc/9RXL-SV7P>) accessed 4th October 2018.

⁹⁵⁰ John Manzoni, ‘Big data in government: the challenges and opportunities’ (*GOVUK*, February 2017) (<https://perma.cc/GF7B-5A2R>) accessed 4th October 2018.

⁹⁵¹ Matthew Woollard, ‘Administrative Data: Problems and Benefits. A perspective from the United Kingdom’ in Adrian Duşa, Dietrich Nelle, Günter Stock and Gert G Wagner (eds), *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences* (SCIVERO Verlag 2014).

⁹⁵² Christopher Hood, *The art of the state: Culture, rhetoric, and public management* (Oxford University Press 2000) 17.

tools usually fused onto existing practices rather than altering them at a deeper level.⁹⁵³ Some scholars have argued that in recent years technologies have taken centre stage, and in doing so have repositioned some of the trajectories of New Public Management into ‘digital era governance’. They point to trends such as the digital re-integration of siloed services, data sharing practices aimed at creating a ‘one-stop-shop’ and ‘end-to-end’ service delivery with minimal repeated information gathering, and in passing mention the rise of interest in ‘zero touch technologies’ (in the terminology of the GDPR and chapters 2 and 3, automated decision-making). As seen, machine learning is thought and hoped to drive a range of these new systems.

I now clarify these terms in this context, distinguishing between two main types of systems using machine learning for operational purposes in the public sector: *automation systems* and *augmentation systems*.

4.1. Automation Systems

Automation systems attempt to increase the quantity or efficiency of routine public sector operations through computation. Here, machine learning is used to enable the automation of tasks which have complicated elements but straightforward and relatively objective outcomes—such as triaging phone-calls or correspondence to the right points of contact.⁹⁵⁴ The incremental automation of rule-based processes is far from new, with public institutions such as tax agencies seeing it as an organisational ambition over many decades, with varying success.⁹⁵⁵ For processes that can be translated to rule-based systems with completeness and fidelity, progress continues at a slow-burn pace.

Many barriers to rote automation surround classic challenges of legacy systems, as well as the slow and surprising creep of information technology in government over time, which has seen a greater fusion of data systems onto locked-in or slow-moving existing practices, rather than the transformative effect that had long been anticipated.⁹⁵⁶ New technologies such as *robotic process automation* have already further aided integration by using computational techniques to automatically connect systems that

⁹⁵³ Helen Margetts, *Information Technology in Government: Britain and America* (Routledge 1999).

⁹⁵⁴ By objective, I am referring to the type of ‘objectivity’ established by methodological approaches such as inter-rater reliability—that different humans making those decisions would arrive at comparably similar results.

⁹⁵⁵ Margetts (n 953).

⁹⁵⁶ cf Anthony Downs, ‘A Realistic Look at the Final Payoffs from Urban Data Systems’ (1967) 27(3) *Public Adm. Rev.* 204 DOI: 10.2307/973283; Patrick Dunleavy, Helen Margetts, Simon Bastow and Jane Tinkler, ‘New Public Management is Dead – Long Live Digital-Era Governance’ (2006) 16(3) *J. Public Adm. Res. Theory* 467 DOI: 10.1093/jopart/mui057.

do not naturally work together.⁹⁵⁷ Similarly, machine learning technologies provide improved tools, such as translation, image or handwriting recognition, which can be ‘plugged in’ to chains of automation for straightforward tasks. This follows the ‘transformative vision’ of information and communication technologies in the public sector, whereby technological innovations can lead to new ‘government instrumentalities and operations’, creating more effective ways of managing public portfolios, and more efficient and personalised public service delivery.⁹⁵⁸

Many administrative tasks are however not straightforward and not easily reduced or defined. Issues concerning operational decision-makers that might *prima facie* appear rote and ‘objective’ may be less so on closer inspection, and instead contain highly subjective and political aspects. Some researchers have historically pointed to a subset of tasks that therefore resist automation. An early empirical study of information systems in US cities concluded that the political nature of some tasks, such as measuring internal departmental goals or deciding on external decisions (eg planning), may never allow them to be dramatically affected by computerisation—and that ‘[p]lanners and policy makers are especially cognizant of this reality’.⁹⁵⁹ ‘Such models’, they argued, ‘would require criteria for defining problems and evaluating solutions, analysis of data in several files, and information that cannot be automated, such as interest group feelings about problems or support for various solutions.’ Given the trend they saw to ‘devalue community statistics and, instead, to emphasize the opinions of the affected citizens’, they claimed ‘it is likely that computerized information will have little impact on city planning decisions in the near future’.⁹⁶⁰ This sentiment has a longer history in public administration, with scholars claiming that ‘the nature of service provision calls for human judgment that cannot be programmed and for which machines cannot substitute’.⁹⁶¹

The implication is that equitable and effective public services require judgement that cannot be quantified, reduced or encoded in fully automated systems. These are issues familiar from the study of artificial intelligence and the law in the early nineties, when it became clear that the application of these systems led to grey zones of knowledge in problem-solving, and that, formally, codification was only effective in ‘some highly specific, syntactically complex but semantically un-troubling domains’.⁹⁶² This

⁹⁵⁷ Leslie P Wilcocks and Mary C Lacity, *Service automation* (Steve Brookes Publishing 2016).

⁹⁵⁸ Christopher C Hood and Helen Z Margetts, *The tools of government in the digital age* (Palgrave Macmillan 2007).

⁹⁵⁹ Alana Northrop, Kenneth L Kraemer, Debora Dunkle and John Leslie King, ‘Payoffs from Computerization: Lessons over Time’ (1990) 50(5) Public Adm. Rev. 505 DOI: 10.2307/976781, 512.

⁹⁶⁰ *ibid* 510.

⁹⁶¹ Michael Lipsky, *Street-level bureaucracy: Dilemmas of the individual in public services* (Russell Sage Foundation 2010) 161.

⁹⁶² Edwards and Veale, ‘Slave to the Algorithm?’ (n 79) 24.

4. Coping with Value(s) in Public Sector Machine Learning

in turn is connected to the indeterminacy of law: particularly the prevalence of terms with an ‘open textured’ nature, where the term’s use or extension cannot be determined in advance of its application;⁹⁶³ where the connections between terms are vague in nature;⁹⁶⁴ or where a series of factors are expected to be weighted and have relative importance assigned in a manner difficult to prescribe or render replicable.⁹⁶⁵ At a larger, more strategic scale, the literature on the governance of sociotechnical problems has similarly emphasised the intractability of ‘unstructured’ or ‘semi-structured’ problems where there is a lack of consensus around appropriate means and/or ends, and how participatory processes that open up rather than close down are required to socially reach more navigable issues.⁹⁶⁶

Automation systems always bring politicised elements in the public sector, from encouraging the shifting and avoidance of blame, the increased rigidity of rules, and the types of ‘edge cases’ on which the systems will fail.⁹⁶⁷ They also serve to prioritise some public values, such as consistency and efficiency, above others.⁹⁶⁸ However, where approaches with significant grey zones are automated, the value-laden nature of automation is accentuated, as the systems have to determine on which basis to make decisions within the grey zones of decision-making. This makes it necessary to ensure that automation systems, particularly ambitious ones, are well-encompassed by frameworks for suitable accountability.

A different perspective to these grey areas also exists, with some arguing that tasks that previously appeared to require human judgement, can now be *better* decided upon with the help of statistical models such as machine learning systems.⁹⁶⁹ This leads to the second category: *augmentation systems*.

4.2. Augmentation Systems

This second category of technological solutions described here, *augmentation systems*, stems from a belief that machine learning does not just help cheapen or hasten decision-making, but can *improve it*.

⁹⁶³ Trevor Bench-Capon and Marek Sergot, ‘Towards a rule-based representation of open texture in law’ in C Walter (ed), *Computer power and legal language* (Quorum Books 1988).

⁹⁶⁴ Henry Prakken, *Logical tools for modelling legal argument* (Kluwer 1997); Zeleznikow (n 277).

⁹⁶⁵ George C Christie, ‘An essay on discretion’ (1986) 5 *Duke Law Journal* 747.

⁹⁶⁶ Hoppe (n 149).

⁹⁶⁷ Matthew L Smith, Merel E Noorman and Aaron K Martin, ‘Automating the public sector and organizing accountabilities’ (2010) 26(1) *Communications of the Association for Information Systems*.

⁹⁶⁸ Christopher Hood, ‘A public management for all seasons?’ (1991) 69 *Public Admin.* 3 DOI: 10/bdwbfj.

⁹⁶⁹ Emma Martinho-Truswell, ‘How AI Could Help the Public Sector’ (*Harvard Business Review*, 26th January 2018) (<https://hbr.org/2018/01/how-ai-could-help-the-public-sector>) accessed 1st November 2018.

What would it be to improve a decision? It is useful to return to a definition of machine learning discussed previously: that a machine learns when its *performance* at a certain *task* improves with *experience*.⁹⁷⁰ Here, performance, task and experience are captured through data, which are determined by designers. Improvement, or learning, can only be discussed once these three areas *at the very least* are formally implemented. At a minimum, this requires that the aims of policy are quantifiable and quantified: a highly value-laden task in and of itself.

Traditionally, ensuring that policy is implemented with fidelity and legitimacy, and that public service delivery decisions are made in an equitable, effective and efficient manner, has fallen within the remit of *bureaucratic professionalism*, which itself carries tensions between responsiveness, as a means of enacting professional judgement, and standardised performance, as a means of ensuring best practice.⁹⁷¹ Bureaucratic professionalism has itself changed from the Weberian model of administrative integrity and impartiality in the public interest, to (new) public management⁹⁷² and there has been growing recognition of its limitations.⁹⁷³ This shift has not only led to increased questioning of the effectiveness of measuring, standardising and auditing public sector performance for the public interest, but also brought about new conceptions of the role of the bureaucrat as negotiator and co-creator of public values with the citizens.⁹⁷⁴ In this respect, one could argue that this shift to new public service (NPS) is supporting the public servant's professional responsibility for more responsiveness in the management and delivery of public services, which augmentation systems could support.

Interestingly, in studies of digitisation of government,⁹⁷⁵ there seems an almost complete omission of anticipation of the augmentative and predictive logics we have seen draw attention today. Programmes such as New Zealand's *Integrated Data Infrastructure* have been designed not (just) for the purpose of creating 'one-stop shops'

⁹⁷⁰ See section 1.3 for a discussion of this definition from Mitchell (n 68).

⁹⁷¹ Richard C Kearney and Chandan Sinha, 'Professionalism and bureaucratic responsiveness: Conflict or compatibility?' (1988) 48(1) Public Adm. Rev. 571; Camilla Stivers, 'The Listening Bureaucrat: Responsiveness in Public Administration' (1994) 54(4) Public Adm. Rev. 364 DOI: 10/dr39pq.

⁹⁷² Carl Dahlström, Victor Lapuente and Jan Teorell, 'The Merit of Meritocratization: Politics, Bureaucracy, and the Institutional Deterrents of Corruption' (2011) 65(3) Polit. Res. Q. 656.

⁹⁷³ G Bevan and C Hood, 'What's measured is what matters: Targets and gaming in the English public health care system' (2006) 84(3) Public Admin. 517 DOI: 10/cww324; Patrick Dunleavy and Christopher Hood, 'From old public administration to new public management' (1994) 14(3) Public Money & Management 9; Christopher Hood and Guy Peters, 'The Middle Aging of New Public Management: Into the Age of Paradox?' (2004) 14(3) J. Public Adm. Res. Theory 267; Irvine Lapsley, 'New Public Management: The Cruellest Invention of the Human Spirit?' (2009) 45(1) Abacus 1.

⁹⁷⁴ Janet V Denhardt and Robert B Denhardt, 'The New Public Service Revisited' (2015) 75(5) Public Adm. Rev. 664; Robert B Denhardt and Janet V Denhardt, 'The New Public Service: Serving Rather than Steering' (2000) 60(6) Public Adm. Rev. 549.

⁹⁷⁵ See eg Patrick Dunleavy, Helen Margetts, Simon Bastow and Jane Tinkler, *Digital Era Governance: IT Corporations, the State and e-Government* (Oxford University Press 2006) DOI: 10/fhj5bz.

4. Coping with Value(s) in Public Sector Machine Learning

for accessing and delivering public services via interoperable, cross-departmental solutions (ie e-Government 1.0 and 2.0), but for the purpose of ‘informing decision-makers to help solve complex issues that affect us all, such as crime and vulnerable children’.⁹⁷⁶ Such programmes are thus established in order to augment the analytic and anticipatory capacity of contemporary governments ‘to systematically use knowledge to inform a more forward-looking and society-changing style of policy-making’.⁹⁷⁷ These augmentations are hoped to help governments navigate coupled and complex problems that have ramifications outside the siloed organisational and decisional structures in which government departments still operate (ie wicked problems).⁹⁷⁸

The nature of the analytic capacity algorithmic augmentation systems are supposed to improve, particularly in the context of linked administrative data combined with additional data sources, is that it is possible to ‘mine’ data for insights public professionals alone would miss. In areas such as tax fraud detection, ambitions do not stay at replicating existing levels of success with reduced staff cost, but to do ‘better than humans’.⁹⁷⁹ In highly value-charged areas where accuracy costs lives, such as child welfare and abuse, it is common to hear calls after a scandal that a tragedy ‘could have been prevented’, or that the ‘information needed to stop this was there’.⁹⁸⁰ Increased accuracy and the avoidance of human bias, rather than just the scalability and cost-efficiency of automation, is cited as a major driver for the development of machine learning models in high stakes spaces such as these.⁹⁸¹

In many ways, this logic continues the more quantified approach to risk and action found in the wide array of managerialist tools and practices associated with New Public Management. These have long had an algorithmic flavour, including performance measures and indicators, targets, and audits. Researchers have also emphasised that while these transformations are often justified as straightforward steps towards greater efficiency and effectiveness, in practice they represent core changes in expectations and in accountability.⁹⁸² Particularly in areas where professional judgement

⁹⁷⁶ Statistics New Zealand (n 949).

⁹⁷⁷ Martin Lodge and Kai Wegrich, *The Problem-solving Capacity of the Modern State: Governance Challenges and Administrative Capacities* (Oxford University Press 2014).

⁹⁷⁸ Leighton Andrews, ‘Public Administration, Public Leadership and the Construction of Public Value in the Age of the Algorithm and ‘Big Data’ Public Admin. DOI: 10/gd25br.

⁹⁷⁹ Cas Milner and Bjarne Berg, *Tax Analytics: Artificial Intelligence and Machine Learning* (PwC Advanced Tax Analytics & Innovation 2017) (<https://perma.cc/4TW3-5P8N>) 15.

⁹⁸⁰ Very rarely do those calling for this consider whether, were systems to be in place using such information, the number of false positives would be small enough to effectively enable identification of particular tragic cases.

⁹⁸¹ Stephanie Cuccaro-Alamin, Regan Foust, Rhema Vaithianathan and Emily Putnam-Hornstein, ‘Risk Assessment and Decision Making in Child Protective Services: Predictive Risk Modeling in Context’ (2017) 79 *Children and Youth Services Review* 291 DOI: 10/gc6c6n.

⁹⁸² Judith Burton and Diane van den Broek, ‘Accountable and Countable: Information Management Sys-

plays a key role in service delivery, such as social work, augmentation tools monitor and structure work to render individuals countable and accountable in new ways, taking organisations to new and more extreme bureaucratic heights of predictability, calculability and control.

* * *

Both public sector automation and augmentation systems based on machine learning clearly risk falling foul of the issues this thesis focusses upon. Firstly, I want to cast a closer look on what those algorithmic issues might look like from the context of public sector practice. I then present the method and results of a study looking to examine how practitioners developing these systems are coping with these challenges on-the-ground.

4.3. Public Values

Work concerning algorithmic issues is useful, but it could be further tailored to the public sector. The need for such tailoring may not seem immediately necessary, but as research has shown, there are significant difference between the values held by the managers of public and private organisations.⁹⁸³ Parallels do exist: companies maximise shareholder value within a strategy set out by the board while civil servants could be argued to be maximising ‘public value’ within a strategy set out by their political masters. Still, what companies seek to maximise is usually better defined (and seemingly more quantifiable) than what civil servants are maximising. As a result, many efforts have made to better understand the landscape of public sector values and put them into some useful framework.⁹⁸⁴

Major discussion also surrounds how public sector values are distinct from ‘public value’, which primarily positions itself as discussing some notion of high quality public management rather than the different ethos and social and ethical aspects that underly it.⁹⁸⁵ Discussions around ‘capacity’ in the public sector⁹⁸⁶ fall into a similar

tems and the Bureaucratization of Social Work’ (2009) 39(7) *The British Journal of Social Work* 1326 DOI: 10/bd5whm.

⁹⁸³ Zeger Van der Wal, Gjalte De Graaf and Karin Lasthuizen, ‘What’s valued most? Similarities and differences between the organizational values of the public and private sector’ (2008) 86(2) *Public Administration* 465 DOI: 10/bwj35.

⁹⁸⁴ The shareholder maximisation theory of the firm is also being increasingly questioned, such as by those emphasising importance of ‘stakeholder value’ in its place. See eg Amy J Hillman and Gerald D Keim, ‘Shareholder value, stakeholder management, and social issues: what’s the bottom line?’ (2001) 22(2) *Strategic Management Journal* 125.

⁹⁸⁵ See eg Mark Harrison Moore, *Creating public value: Strategic management in government* (Harvard University Press 1995).

⁹⁸⁶ See eg Lodge and Wegrich (n 977).

4. Coping with Value(s) in Public Sector Machine Learning

Value	Description
Reliability	Stability and adaptability in the face of outside influences
Equity	Fair and impartial treatment
Accountability	Requiring scrutiny and justification of action with a view to blame, punishment or redress
Legality	Being in accordance with the law
Dialogue	Broad, responsive and two-way stakeholder involvement
Usability	Designing service delivery with consideration of diverse human factors
Innovation	Creating or applying novelty
Upskilling	A working environment promoting self-development
Productivity	Efficiency and effectiveness
Competition– Cooperation	Utilising the incentives and information cues of markets versus building institutions to solve collective action problems
Openness– Secrecy	Balancing publishing what is done with privacy and strategic confidentiality
Advocacy– Neutrality	Defending interests (e.g. the environment) versus professional impartiality

Table 4.1.: Public sector values relevant to machine learning.

trap of struggling to define the ‘good’, or assuming perhaps that it is a matter of due process or even ‘common sense’. But these components are slippery, wide-ranging, and contentious. There can be no exhaustive list that categorises ethical concerns, nor can there be one objective typology or hierarchy.

While objective frameworks in this value-charged area are by definition impossible, this author feel that serviceable ones can exist, and can be useful. One classic paper that claims to make a sensible yet parsimonious framework for public sector values⁹⁸⁷ distinguishes between three sets of core values in public management. In the original work, these values are somewhat unhelpfully termed *sigma*, *theta* and *lambda*. *Sigma* values focus on keeping the public sector ‘lean and purposeful’, *theta* values focus on keeping it ‘honest and fair’, and *lambda* values focus on keeping it ‘robust and resilient’.

Other scholars have generated more expansive lists. One review examines 420 papers on public values to attempt to reach some common ‘constellations’,⁹⁸⁸ arriving

⁹⁸⁷ Hood, ‘A public management for all seasons?’ (n 968).

⁹⁸⁸ Torben Beck Jørgensen and Barry Bozeman, ‘Public values: An inventory’ (2007) 39(3) Administration

at a lengthier list. A subset of these are shown in Table 4.1. While the original list included broader values such as regime stability and majority rule, here it is narrowed to only those values that tally with the criteria of this study. While no value is uncontested, some are more contested than others. Some values or nuances of values are the material that elections are fought over. Other values are relatively less contentious, although may be geographically specific. In the UK, features such as impartiality and stability of the civil service, due process, the rule of law and the upholding of constitutional conventions are among aspects that have historically seemed very rarely contested.⁹⁸⁹ In other places, the composition may be different, but such a spectrum tends to exist. Identifying some of the values that are highly likely to vary and be the subject of decisions in particular projects or tasks allows for a narrowing down of the areas concerned in the investigation to come.

This is not to state that these values could not or even should not be otherwise, suffice to say that they are rarely put to vote or politicised. Deeper issues around the role of algorithmic systems, such as their connection to notions of justice in general,⁹⁹⁰ must be discussed. Being concerned around equity or robustness of algorithmic systems within high-stakes, public sector settings is no substitute from considering how the aggregate actions and approach of the public sector as a whole to the quantification of citizens might have undesirable effects.⁹⁹¹ Yet equally, a failure to understand how these issues play out *within* applications, such as particular instances of software systems of policy initiatives, could cause similar harm in practice. In short, both narrower conceptions of application and values as well as broader considerations of the appropriateness of machine learning tools in certain settings and policy areas are needed. Luckily, scholars are not faced with false dichotomies and mutually exclusive choices in practice, and the research community can do more than one thing.

As a consequence, the values broadly considered, and used to steer and arrange the findings in the study to follow are those the review mentioned above⁹⁹² described as as connecting the public sector to its *environment*, its *employees*, its *citizens* or *itself*. These are listed in Table 4.1.

& Society 354 DOI: 10/ch6kwwg.

⁹⁸⁹ Particularly given the times in which this thesis is being written, the author hopes this remains the case, both in the UK and across the world.

⁹⁹⁰ See eg Jeffrey Alan Johnson, 'From Open Data to Information Justice' (2014) 16(4) *Ethics and Information Technology* 263 DOI: 10/gfgt36; Linnet Taylor, 'What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally' (2017) 4(2) *Big Data & Society* 2053951717736335 DOI: 10/gfgt4b.

⁹⁹¹ See also Scott (n 34).

⁹⁹² Jørgensen and Bozeman (n 988).

4.4. Questions

Public sector values serve as a point of departure for the following examination of high-stakes machine learning on the ground. There are several reasons for this.

Firstly, without sociotechnical understanding of how, operationally, to meet the challenge of core public sector values in machine learning, it seems somewhat difficult to contest deeper issues. Insofar as procedurally, they are a hard-earned pre-requisite for ‘good governance’, they seem quite desirable across ideological lines. They are no panacea to deeply entrenched inequalities, nor for political fights, but politics of many colours are likely to need to know at minimum how to meet them in the systems they deploy.

Public sector values, in some ways, are broader than the values discussed in the machine learning ethics literature. Issues like robustness, reliability and resilience are rarely raised, for example—even though the robustness of machine learning is a research topic within computer science literature, and robustness of public sector organisations is a core feature of stable societies. These concerns around a disconnect between technical governance and on-the-ground needs result in two core research questions that drive the *in vivo* machine learning research in this thesis. The existing literature has been drawn from a base that has to date largely ignored on-the-ground challenges of practitioners through both lack of evidence and lack of co-design.⁹⁹³ As a result, there is real concern that the tools that have been assumed as useful in communities such as FATML do not map to practitioner experiences and needs, and therefore will struggle to garner significant adoption.

As discussed in section 1.7, the main question this chapter seeks to answer is: how congruent are the assumptions and framings of contemporary computational tools designed to tackle social issues in algorithmic systems with real-world environments and constraints? It is worth unpacking this. Firstly, congruence in this context accepts, to a certain degree, the framings that public sector organisations whose interviews this chapter reports upon place around these technologies. These might not be the ‘right’ frames when considered from a variety of societal perspectives. Some pressure groups have argued that even when a public sector machine learning system seems to perform in ‘less biased’ ways than humans, there is a moral ground not to utilise them.⁹⁹⁴ I do not argue that these frames are ‘right’ or that they should be accepted uncritically, but I *do* argue they need to be taken seriously, as they often have a grounding in perceptions and problem structures that can go unobserved when ana-

⁹⁹³ See also Holstein, Wortman Vaughan, Daumé, Dudik and Wallach (n 15).

⁹⁹⁴ See Liberty, ‘Liberty’s written evidence to the Select Committee on Artificial Intelligence’ (*House of Lords Select Committee on Artificial Intelligence*, 1st September 2017) (<https://perma.cc/9LAQ-JWWM>) para 23.

lysing a problem from afar. In particular, understanding these frames also provides insight on the institutions that have emerged around them, which is useful knowledge for intervention. As systems researcher Donella Meadows argued⁹⁹⁵

Before you disturb the system in any way, watch how it behaves [...] Ask people who've been around a long time to tell you what has happened. [...] Starting with the behavior of the system forces you to focus on facts, not theories. It keeps you from falling too quickly into your own beliefs or misconceptions, or those of others. [...] Starting with the behavior of the system directs one's thoughts to dynamic, not static analysis—not only to 'what's wrong?' but also to 'how did we get there?' and 'what behavior modes are possible' and 'if we don't change direction, where are we going to end up?' [...] [S]tarting with history discourages the common and distracting tendency we all have to define a problem not by the system's actual behavior, but by the lack of our favorite solution.

The focus on the public sector in attempting to ask this question is strategic from a number of angles. Firstly, public sector organisations often make consequential decisions, and as a result, their machine learning is likely to 'matter'. Secondly, the fact that transparency has or is a public value enables access to actors that might be difficult to achieve in a private sector context, particularly in competitive industries that value trade secrecy. Such transparency should in theory even be compounded by how the subject of interest is the execution of public values in themselves. Authors since the first published version of this study⁹⁹⁶ (with industry access aided by their membership of an industrial lab) have undertaken a partial replication of it for the private sector, and so it does appear that both contexts are amenable for this type of research.⁹⁹⁷ Keeping them distinct does usefully serve a third aim: to examine machine learning development strategies in specific light of the public administration literature on public values already discussed.

4.5. Method

27 individuals agreed to be interviewed in late 2016—predominantly public servants and attached contractors either in modelling or project management. Each individual

⁹⁹⁵ Donella Meadows, 'Dancing with systems' (2002) 13(2) *The Systems Thinker* (<https://perma.cc/4RPZ-XN3H>).

⁹⁹⁶ Michael Veale, Max Van Kleek and Reuben Binns, 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'18)* (ACM 2018) DOI: 10/ct4s.

⁹⁹⁷ Holstein, Wortman Vaughan, Daumé, Dudik and Wallach (n 15).

4. Coping with Value(s) in Public Sector Machine Learning

was interviewed only once, either in person (17 interviews) or on the telephone (10 interviews). They were all undertaken with only one interviewer—the author—and each lasted between 40–60 minutes. Just over a fifth of informants were female. Interviewees worked in one of 5 OECD countries located over 3 continents. It was decided in the initial ethical approval for this work not to publicly name the countries in order to reduce the risk of informant identification, particularly by the uniqueness of country added to role, but I will note, given that it has been the source of many algorithmic war-stories,⁹⁹⁸ that the US was *not* one of the countries included.

Informants were identified with an *ad hoc* sampling approach. This was chosen for several reasons. Firstly, at this relatively early stage of deployment, projects are emerging without central mandates—no coordinating body was identified to have a reliable compiled register of activities. Indeed central agencies occasionally shared registers that turned out to be a poor representation of on-the-ground activities.⁹⁹⁹ Secondly, I sought as many perspectives as possible from within public sector organisations deploying machine learning for decision-support today, and felt this was best achieved by looking across sectors to very different types of agencies and bodies. To recruit participants, projects and contacts were assembled from grey literature, freedom of information requests (both actively made and through the historical requests on platforms such as *WhatDoTheyKnow*¹⁰⁰⁰), snowball sampling, direct inquiries with organisational contacts, and the use of news databases including *Factiva* and *Lexis-Nexis*. Terms including *predictive modelling*, *entity-level prediction*, *predictive analytics*, *big data* and *machine learning* were entered into these databases and public document repositories. Participants additionally played an important role in sampling themselves, and were usually willing and often even eager to flag colleagues in other domestic or foreign agencies working on projects they felt would benefit the study. Similarly to challenges arranging interviews with societal ‘elites’, candidacy for interviews ‘often cannot be planned for adequately in advance of the project; rather, it emerges as part of the fieldwork’.¹⁰⁰¹

Because of the open-ended nature of the sampling, the varied nature of the roles (particularly across sectors), and the many different systems concerned, it was neither possible nor helpful to stick to a rigid interview script. Instead, the approach taken was similar to other open-ended work in policy research, involving prompting the parti-

⁹⁹⁸ See section 1.4: in particular the US-based war-stories concerning Recidivism and Racism, Discrimination and Search Engines and Know Your Customers.

⁹⁹⁹ The author noted this at an early stage in relation to a central UK government project on machine learning in policy-making he was involved in.

¹⁰⁰⁰ <https://www.whatdotheyknow.com/>.

¹⁰⁰¹ Teresa Odendahl and Aileen M Shaw, ‘Interviewing elites’ in Jaber F Gubrium and James A Holstein (eds), *Handbook of Interview Research* (SAGE 2002) DOI: 10/gfgq8h.

participant to not only outline their role but explain the process behind the development and maintenance of the project.¹⁰⁰² First, the purpose of the study was explained to participants, at which point any ambiguities could be resolved. Following that, participants were asked about their role (and history of roles) in this area, then to give a high level outline of relevant project(s) and a more detailed view on their position within them. They were then steered at opportune moments in the discussion towards topics of fairness and accountability, effectiveness and complexity or robustness, mirroring the the public sector values framework introduced by Hood and discussed above.¹⁰⁰³ At times, this steering was unnecessary and avoided, particularly as the nature of the study was made clear to participants: many already had considered these issues during their job, albeit often under different names.

The other main prompt used to elicit relevant insights, particularly where participants had not considered their job in the above framing before, was to ask whether ‘anything surprising or unexpected’ had happened to them in relation to their work, such as a deployed model.¹⁰⁰⁴ This was especially useful in eliciting institutional events, or novel incidences of model failure, as well as widening the discussion to beyond the narrow topics which the author had arrived with preconceptions of. It has been adopted since by at least one study seeking to replicate and extend the findings described below.¹⁰⁰⁵

Conversations were not taped. While that might have been desirable, recording audio of individuals discussing sensitive public sector work is extremely difficult. Bodies commonly disallow it when individuals are not spokespersons for the organisation, precluding it as an interview approach, more so where new technologies are involved, and questions asked are likely to be totally new. Even where taping is permitted, it can risk inhibiting openness and frankness in discussions. These politically-charged contexts pose methodological restrictions infrequently seen in areas such as human–computer interaction,¹⁰⁰⁶ but frequently encountered by public administration researchers, and I therefore followed methodological practices developed in the latter field.¹⁰⁰⁷ These are further exacerbated here by fear of negative media coverage—both journalists and academics in this field have exhibited a recent taste for the ‘shock

¹⁰⁰² See eg Edward C Page and Bill Jenkins, *Policy bureaucracy: Government with a cast of thousands* (Oxford University Press 2005), who conducted 128 interviews in the UK civil service to understand the nature of policy work, asking only ‘what do you do?’ and ‘how do you come to be in this job?’.

¹⁰⁰³ See Hood, ‘A public management for all seasons?’ (n 968) and section 4.3.

¹⁰⁰⁴ This is, incidentally, also the interviewing method of choice in *Private Eye*’s satirical column ‘Me and My Spoon’, which pretends to have asked celebrities inane cutlery-related questions which always end in asking if ‘anything amusing ever happened to you in connection with a spoon?’.

¹⁰⁰⁵ Holstein, Wortman Vaughan, Daumé, Dudik and Wallach (n 15).

¹⁰⁰⁶ Where an earlier version of this chapter was published as Veale, Van Kleek and Binns (n 996).

¹⁰⁰⁷ See eg Page and Jenkins (n 1002).

stories' of which many of the previously discussed algorithmic war-stories serve as examples of.¹⁰⁰⁸ Instead, verbose notes were continuously taken with the aim of authentically capturing both interviewees' tone, phrasing and terminology, as well as the core points they explained. Where longer continuous answers were given, interviewees kindly paused for note-taking purposes. Notes were typed up by the author, always on the same day as the interview took place and often immediately following. Some highly context-specific terminologies, such as geographic subunits or revealing were substituted with approximately equivalent generic alternatives to increase the difficulty of project re-identification. Handwritten notes were then destroyed in line with data protection and the study's ethical approval.

To analyse the interviews, open coding was used (utilising *NVivo 11 for Mac*), with codes iteratively generated and grouped concerning the challenges and coping mechanisms observed. These were then iteratively grouped according to a public sector values framework from the public administration literature¹⁰⁰⁹ as a thematic organisational principle. Given that the coding was for organisational purposes rather than for final analytic communication or quantification, as well as in light of the sensitivity of the data in question, multiple raters and interrater reliability metrics were not used.

4.6. Findings

Below I detail the findings from the work in a narrative grouped by two major headings internal and external actors. This structure was selected upon examining the structured data for flow and clarity rather than theoretical basis. It does however emphasise the importance of the systems lens already described in this thesis, as well as helping challenging preconceptions of systems boundaries implicitly drawn by FATML tools such as those outlined in section 1.6 by emphasising dynamics that seem important to practitioners but are often not currently considered 'in scope' by computer scientists.

4.6.1. Internal actors and high-stakes machine learning

The first category of themes relate to discussions of how the deployed systems are connected to and perceived by a range of internal actors. Many public concerns around algorithmic transparency have so far, as already discussed, focussed on external algorithmic accountability and transparency-based rights, such as a 'right to an explan-

¹⁰⁰⁸ See section 1.4.

¹⁰⁰⁹ Jørgensen and Bozeman (n 988).

ation’,¹⁰¹⁰ although broad reasons to make systems transparent and interpretable exist.¹⁰¹¹ Within organisations informants displayed a wide array of reasons for understanding data and models, and their responses demonstrated how broad these reasons could be in practice.

4.6.1.1. Getting individual and organisational buy-in

Informants reported a need to use different approaches to clarify the workings of or process behind machine learning powered decision-support systems for internal actors. Some of these were strategic actors in management positions, either the clients of external contractors or customers of internal modelling teams.

Several interviewed practitioners noted that this organisational pressure led them to make more ‘transparent’ machine learning systems. Detection systems for fraudulent tax returns illustrated this. The analytics lead at one tax agency [X1] noted that they ‘have better buy-in’ when they provide the logic of their machine learning systems to internal customers, while their counterpart in another tax agency [X2] described a need to ‘explain what was done to the business user’. Both these individuals and modellers around them emphasised they had in-house capability for more complex machine learning systems, such as support vector machines or neural networks, but often chose against them for these reasons. Instead, many of the systems that ended up being deployed were logistic regression or random forest based.

Some saw transparency in relation to input variables more than model family. One contractor that constructed a random-forest based risk score for gang members around knife crime on behalf of a global city’s police department [X3] described an ‘Occam’s razor’ process, where they started with 18,000 variables, working down to 200, then 20, then 8—‘because it’s important to see how it works, we believe’. To steer this, they established a target percentage of accuracy with the police department *before* modelling—around 75%—which they argued helped them avoid trading off transparency. When users of analytics are not ‘confident they know what a model is doing’, they ‘get wary of picking up protected characteristics’, noted the modelling lead at tax agency [X4]. To make this more transparent, the police contractor above [X3] would ‘make a model with and without the sensitive variables and see what lift you get in comparison’, presenting those options to the client to decide what was appropriate.

Another issue raised by several modellers was the difficulty in communicating the performance of designed systems. One modeller in a regional police department [X5]

¹⁰¹⁰ See section 2.2.2.

¹⁰¹¹ See section 1.6.1.

4. Coping with Value(s) in Public Sector Machine Learning

was designing a collaborative system with neighbouring police departments to anticipate the location of car accidents. They noted that

We have a huge accuracy in our collision risk, but that's also because we have 40 million records and thankfully very few of them crash, so it looks like we have 100% accuracy—which to the senior managers looks great, but really we only have 20% precision. The only kind of communication I think people really want or get is if you say there is a one-fifth chance of an accident here tomorrow—that, they understand.

An analytics lead at a tax department [X2] faced parallel issues. When discussing the effectiveness of a model with clients, he would often find that 'people tend to lose faith if their personally preferred risk indicators aren't in a model, even without looking at performance of results.'

Performance was often judged by the commissioning departments or users based on the additional insight it was thought to provide, compared to what they thought to be known or easily knowable. There was a tension between those who were seeking insight beyond existing processes, and those seeking efficiency/partial automation of current processes. One contracted modeller for a police department [X3] noted that during modelling, they 'focussed on additionality. The core challenge from [the police department] was to ascertain whether the information we could provide would tell them things they did not already know. How would it complement the current way of doing things?' Yet another case, an in-house police modeller [X6] noted that a focus on additionality by the users of the system often clouded the intended purpose of the software in the first place.

What we noticed is that the maps were often disappointing to those involved. They often looked at them and thought they looked similar to the maps that they were drawing up before with analysts. However, that's also not quite the point—the maps we were making were automatic, so we were saving several days of time.

4.6.1.2. Over-reliance, under-reliance and discretion

Over and under-reliance on decision support, extensively highlighted in the literature on *automation bias*,¹⁰¹² featured considerably in informants' responses. A lead machine learning modeller in a national justice ministry [X7], whose work allocates

¹⁰¹² Skitka, Mosier and Burdick (n 232); Dzindolet, Peterson, Pomranky, Pierce and Beck (n 361).

resources such as courses within prisons, described how linking systems with professional judgement ‘can also mean that [the model output is] only used when it aligns with the intuition of the user of the system’. To avoid this, some informants considered more explicitly how to bring discretion into decision-support design. A lead of a geo-spatial predictive policing project in a world city [X8] noted that they designed a user interface

to actively hedge against [officers resenting being told what to do by models] by letting them look at the predictions and use their own intuition. They might see the top 3 and think ‘I think the third is the most likely’ and that’s okay, that’s good. We want to give them options and empower them to re-view them, the uptake will hopefully then be better than when us propellor-heads and academics tell them what to do...

Model outputs were not treated similarly as decision support in all areas. The former lead of a national predictive policing strategy [X9] explained how they saw discretion vary by domain.

We [use machine learning to] give guidance to helicopter pilots, best position them to to optimise revenue—which means they need to follow directions. They lose a lot of flexibility, which made them reluctant to use this system, as they’re used to deciding themselves whether to go left or right, not to be told ‘go left’! But it’s different every time. There were cases where agents were happy to follow directions. Our police on motorcycles provide an example of this. They were presented with sequential high risk areas where criminals should be and would go and apprehend one after another—and said “yes, this is why we joined, this is what we like to be doing!” The helicopters on the other hand did not like this as much.

Also faced with a list of sequential high risk activities, this time relating to vulnerability of victims, the analytics lead at one regional police department [X10], sought advice from their internal ethics committee on how to use the prioritised lists their model outputted.

We had guidance from the ethics committee on [how to ethically use rank-ordered lists to inform decision-making]. We were to work down the list, allocating resources in that order, and that’s the way they told us would be the most ethical way to use them [...] It’s also important to make clear that the professional judgement always overrides the system. It is just another tool that they can use to help them come to decisions.

4.6.1.3. Augmenting models with additional knowledge

Informants often recognised the limitations of modelling, and were concerned with improving the decisions that were being made with external or qualitative information. A lead of a national geospatial predictive policing project [X11] discussed transparency in more social terms, surrounding how the intelligence officers, who used to spend their time making patrol maps, now spent their time augmenting them.

We ask local intelligence officers, the people who read all the local news, reports made and other sources of information, to look at the regions of the [predictive project name] maps which have high predictions of crimes. They might say they know something about the offender for a string of burglaries, or that building is no longer at such high risk of burglary because the local government just arranged all the locks to be changed. [...] We also have weekly meetings with all the officers, leadership, management, patrol and so on, with the intelligence officers at the core. There, he or she presents what they think is going on, and what should be done about it.

Other types of knowledge that modellers wished to integrate were not always fully external to the data being used. In particular, information needs also arose linked to the primary collectors of training data. One in-house modeller in a regional police department [X5], building several machine learning models including one to predict human trafficking hotspots, described how without better communication of the ways the models deployed worked, they risked large failure.

Thankfully we barely have any reports of human trafficking. But someone at intel got a tip-off and looked into cases at car washes, because we hadn't really investigated those much.¹⁰¹³ But now when we try to model human trafficking we only see human trafficking being predicted at car washes, which suddenly seem very high risk. So because of increased intel we've essentially produced models that tell us where car washes are. This kind of loop is hard to explain to those higher up.

Similarly, external factors such as legal changes can present challenges to robust modelling. A modeller in a justice ministry building recidivism prediction systems [X7] noted that while changes in the justice system were slow, they were still 'susceptible to changes in sentencing, which create influxes of different sorts of people into the prison systems.' These kinds of rule change are unavoidable in a democratic society,

¹⁰¹³ Car washes are a major location where modern day slavery is found. See eg BBC News, 'Kent slavery raids 'uncover 21 victims'' (*BBC News*, 6th December 2016) (<https://perma.cc/AM4S-RMHR>).

but awareness of them and adequate communication and preparation for them is far from straightforward.

4.6.1.4. Gaming by decision-support users

‘Gaming’ or manipulation of data-driven systems, and the concern of this occurring if greater transparency is introduced, is often raised as an issue in relation to the targets of algorithmic decisions. This will shortly be discussed in a subsequent theme. Yet types of *internal gaming* within organisations have received considerably less treatment by those concerned about value-laden challenges around algorithmically informed decisions. This is despite how internal gaming is extensively highlighted in the public administration literature in relation to targets and the rise of New Public Management,¹⁰¹⁴ a broad movement towards ‘rationalisation’ in the public sector that clearly affected informants around the world.

One tax analytics lead [X2] worried that releasing the input variables and their weightings in a model could make their own auditors investigate according to their perception of the model structure, rather than the actual model outputs—where they believed that bias, through fairness analysis,¹⁰¹⁵ could ostensibly be controlled.

To explain these models we talk about the target parameter and the population, rather than the explanation of individuals. The target parameter is what we are trying to find—the development of debts, bankruptcy in six months. The target population is what we are looking for: for example, businesses with minor problems. We only give the auditors [these], not an individual risk profile or risk indicators [...] in case they investigate according to them.

Additionally, some tax auditors are tasked with using the decision-support from machine learning systems to inform their fraud investigations. Yet at the same time, the fraud they discover feeds future modelling; they are both decision arbiter and data collector. The effect these conflicting incentives might have on a model were highlighted by a different tax agency [X2], as when auditors accumulate their own wages, ‘[i]f I found an initial [case of fraud], I might want to wait for found individuals to accumulate it, which would create perverse incentives for action.’

¹⁰¹⁴ Bevan and Hood (n 973).

¹⁰¹⁵ See section 1.6.2, p. 65.

4.6.2. External actors and high-stakes machine learning

This second theme focusses on when informants reflected upon value concerns that related to both institutional actors that were outside their immediate projects, or that were at a distance, such as subjects of algorithmically informed decisions.

4.6.2.1. Sharing models and pushing practices

Scaling-up is an important part of experimentation. This is particularly the case in public sector organisations replicated by region—while some of them, particularly those in the richest or densest areas, can afford to try new, risky ideas with the hope of significant performance or efficiency payoffs to outweigh their investment, for smaller or poorer organisations that economic logic does not balance. The latter set of organisations are more reliant on the import and adaptation of ideas and practices from more well-resourced sister organisations (which could also be abroad) or from firms. Yet in practice, this is challenging, as machine learning systems also come imbued with very context specific assumptions, both in terms of the problem they are attempting to model, and the expertise that surrounds the decision-making process each day it is used. A modeller and software developer in a spatiotemporal predictive policing project [X6] emphasised the challenges in scaling up these social practices, as they were not as mobile as the software itself.

If you want to roll out to more precincts, they have to actually invest in the working process to transform the models into police patrols. To get more complete deployment advice... it takes a lot of effort to get people to do that. What you see is that other precincts usually—well, sometimes—set up some process but sometimes it is too pragmatic. What I mean by this is that the role of those looking at the maps before passing them to the planner might be fulfilled by someone not quite qualified enough to do that.

Similar sentiments were also echoed by individuals in national tax offices, particularly around the ‘trading’ of models by large vendors. One tax analytics lead [X2] in a European country expressed concerns that another less resourced European country was being sold models pre-trained in other jurisdictions by a large predictive analytics supplier, and that they would not only transpose badly onto unique national problems, but that the country interested in purchasing this model seemed unprepared to invest in the in-house modelling capacity needed to understand the model or to change or augment it for appropriate use.

4.6.2.2. Accountability to decision subjects

Interpretable models were seen as useful in relation to citizens. One lead tax analyst [X2] described how transparency provided ‘value-add, particularly where an administrative decision needs explaining to a customer, or goes to tribunal.’ They noted that ‘sometimes [they] justifi[ed] things by saying here are the inputs, here are the outputs’ but they were ‘not really happy with that as an ongoing strategy.’ Yet on occasion, more detailed explanations were needed. The same informant recalled an incident where a new model, in line with the law, was flagging tax deductions to refuse that were often erroneously allowed to some individuals in previous years. Naturally, many people called in to complain that their returns were not processed as expected—so the tax agency had to build a tool to provide call centre operators with client-specific explanations.¹⁰¹⁶

Other organisations focussed on providing knowledge of the system to other interested parties, such as media organisations. One national predictive policing lead [X11] explained how they found it difficult to have discussions around equity and accountability with police officers themselves, who are often narrowly focussed on ‘where they think they can catch someone’, and have less capacity or incentive to devote time and energy to frame broader questions. Instead, this police force would invite journalists over twice a year to see what the predictive teams ‘do, how [the algorithms] work, and what we are doing.’ Several public sector organisations using machine learning systems already publish information about the weights within their model, the variable importance scores, or record ethical experiences and challenges in the modelling process.¹⁰¹⁷

4.6.2.3. Discriminating between decision subjects

Discrimination has taken centre-stage as the algorithmic issue that perhaps most concerns the media and the public. Direct use of illegal-to-use protected characteristics was unsurprisingly not found, and interviewees were broadly wary of directly using protected characteristics in their models. Input data was seen as a key, if not the only point of control, but the reasons and the logics behind this varied. A lead of analyt-

¹⁰¹⁶ It was unclear whether this model was machine-learning based, or just a rule-based system.

¹⁰¹⁷ See eg Nikolaj Tollenaar, BSJ Wartna, PGM Van Der Heijden and Stefan Bogaerts, ‘StatRec — Performance, validation and preservability of a static risk prediction instrument’ (2016) 129(1) *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 25 DOI: 10/gfgrbb; Robin Moore (ed), *A compendium of research and analysis on the Offender Assessment System* (Ministry of Justice Analytical Series 2015) (<https://perma.cc/W2FT-NFWZ>); Marion Oswald, Jamie Grace, Sheena Urwin and Geoffrey C Barnes, ‘Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and ‘Experimental’ Proportionality’ (2018) 27(2) *Information & Communications Technology Law* 223 DOI: 10/gdz288.

4. Coping with Value(s) in Public Sector Machine Learning

ics at a national tax agency [X2] noted that ‘if someone wanted to use gender, or age, or ethnicity or sexual preference into a model, [they] would not allow that—it’s grounded in constitutional law.’ In one case guidance was to be released clarifying forbidden variables, but made no difference as the tax agency was already compliant [X1]. Even when characteristics were found to be legally permitted after consultation with lawyers (characteristics are not protected in all contexts), they might still have been avoided. Informant [X4], a lead modeller in a tax agency, noted that they have an informal list ‘of variables that [they] don’t feed into models’, which included age and location, both of which were legally permissible in their context. Location was avoided by this informant because even though different cities have different tax fraud risks, they ‘don’t usually want to investigate on those grounds.’ In other cases, home location was avoided as it was a ‘proxy for social deprivation’, in the words of the lead modelling a justice ministry [X7].

Occasionally, there would be pressure to use protected characteristics to increase predictive power. The same justice modeller [X7], noted that ‘we had feedback from a senior [foreign nationality, omitted] academic in this space on our [criminal justice] model, noting that ‘if you’ve got something as predictive as race is, why aren’t you using it?’ Many of [this experts’ deployed] models do, but it’s an ethical decision in my mind and this is the route we’ve taken.’ In this vein, they were also concerned with the interaction between sensitive variables and the proxy outcome that could be measured (conviction) rather than the outcome variable of true interest (offending).

Race is very predictive of re-offending, [but] we don’t include race in our predictive models [...] we are aware that we are using conviction as the proxy variable for offending, and if you do this then you can get into cycles looking at certain races which might have a higher chance of being convicted, and train models on this data instead. That would mean you’re building systems and catching people not based on the outcome, but on proxy outcomes.

Lastly, it can be highlighted that technical methods for debiasing were often not those relied on by practitioners, who instead had to draw on existing research to describe the kind of correlations that might occur. A police officer working on a predictive child protection system [X12] noted that ‘whether a child is deaf or disabled is empirically linked to abuse, according to [civil society organisation] research’, using this to justify the type of data they were requesting and concerns they were raising during the mode-building process.

4.6.2.4. Gaming by decision subjects

It is commonly expressed that extensive transparency of algorithms to the public might encourage system gaming,¹⁰¹⁸ and this is brought to bear as a justification for opacity. Correspondingly, external gaming was raised as an issue by some informants. One contractor developing predictive policing software for a world city [X3] noted that concerns in his sector concerned ‘criminal gangs that might send nine guinea pigs through the application process looking for loopholes to get arrested, just to find a tenth that highlights a way they can reliably get passports from under the noses of the authorities.’ An analyst from a large NGO working in collaboration with the police on developing a predictive system to detect child abuse [X13] noted that ‘it’s much harder to game when you’ve linked up lots of different aspects, education and the like’, although their colleague [X14] warned that they were concerned about many of the usual sophisticated practices being used to game ML-supported systems, such as ‘turning professionals against each other’ or the ‘strategic withholding of consent at opportune moments.’ The analytics lead at one tax agency [X1] explained that while they would publicly share the areas they were interested in modelling tax fraud for, such as sectors or size, they were ‘primarily concerned that if the model weights were public, their usefulness might diminish.’

Other incidents resembled gaming—and could feasibly be interpreted as such—but served more to demonstrate the current fragility of models towards concerted attempts to change them. A modeller at a police department [X5] noted, in relation to a model they had built to pre-empt when the force should ensure they had the most staff available to deal with missing persons, that

There’s one woman who calls in whenever her kid is out after 10pm. She then calls back about 30 minutes or so later to say that everything is fine, or we follow up with her. But then it looks like in the model that kids always go missing at 10pm, which obviously is a bit misleading. In the end I had to manually remove her from the model to remove the spurious pattern.

While in this case, the model failed—resembling an *availability attack*, to draw on terms from cybersecurity literatures relating to the availability of service—this might not always be the case. Indeed, models might not fail in obvious ways, or might even be subject to attacks designed to change them in targeted ways.¹⁰¹⁹ Even where an attack is not planned, simply responding to decisions informed by the model—such as patrol

¹⁰¹⁸ See eg ‘More accountability for big-data algorithms’ (2016) 537(7621) *Nature* 449 DOI: 10/gfgrbj.

¹⁰¹⁹ See generally Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin I P Rubinstein and J D Tygar, ‘Adversarial machine learning’ in *Proceedings of the 4th ACM workshop on Security and Artificial Intelligence* (2011) DOI: 10/ft95kn.

patterns—might look like gaming, or at least a game of cat-and-mouse. The police lead on a geospatial predictive policing project for a world city [X8] noted this in their own system. While it wasn't clear whether they were just removing the lowest hanging fruit or criminals were responding, in response, they linked a further feedback effect to try to compensate for the performance loss.

The highest probability assessments are on the mark, but actual deployment causes displacement, dispersion and diffusion, and that throws the algorithm into a loop. You have to remodel, though typical patterns of unresponded-to crime are predicted well [...] we decided to re-evaluate learning every 2–3 weeks, pull in all sorts of other variables, such as feeding it with what police were deployed, what they did—I've never seen this in other similar systems. In the first four weeks of trialling it out, the probability of being correct just tanked [...] in the third update, it started to figure shit out.

4.7. Implications for research and practice

I now draw together some overarching findings from the previous sections, representing areas I consider as phenomena or themes under-emphasised or somewhat at tension with the types of FATML research described in section 1.6.

4.7.1. 'The probability of being correct tanked': Data changes

Data in the public sector is usually collected, categorised and cleaned for primarily operational reasons, such as recording who has been arrested, calculating tax bills, or delivering mail—not for modelling. While increasing emphasis is now put on secondary uses of data,¹⁰²⁰ primary needs remain primary. Upstream changes in the logic of collection—as was the case above when investigative patterns led to a human trafficking risk model becoming a car-wash detector¹⁰²¹—can have significant downstream effect. Particularly where models are quietly being developed or piloted, or are located in a different part of the organisation from data collection efforts, it is easy for changes in practices to occur without those responsible for model performance to be aware of them. Accountability becomes difficult to trace in these situations. As Nick

¹⁰²⁰ Administrative Data Taskforce, *The UK Administrative Data Research Network: Improving access for research and policy* (Economic and Social Research Council 2012) (<http://www.esrc.ac.uk/files/research/administrative-data-taskforce-adt/improving-access-for-research-and-policy/>).

¹⁰²¹ See above at section 4.6.1.3, p. 240.

Seaver puts it, these systems are ‘not standalone little boxes, but massive, networked ones with hundreds of hands reaching into them’.¹⁰²² Accountable systems should to be internally accountable, else it would appear to be difficult for external accountability to either make sense or be sustained.

Data can also change *because* of the model rather than in spite of it. Where model results determine the behaviour of the same resources that also collect data then they are directly influencing the future sampling of their training data.¹⁰²³ Sending police officers to areas of high predicted crime is an example of this. In the worst cases, the model can have a polarising effect: directing police resources disproportionately to areas with slightly higher crime risk will, without corrections, skew future training data collection in those area, which might be demographically or socioeconomically disproportionate.¹⁰²⁴ In other cases, effects might be more subtle but of equal importance, and might cause particular failures to occur in unforeseen ways. If individuals react to try and influence or game a system—as the example stories above indicate is certainly possible—then the future population distribution becomes a function of past model decisions or structure. Little work has focussed on this so far, particularly on the impacts of these on the research into statistical fairness and non-discrimination properties, which broadly implicitly assume stationarity in their problem set-up. This is also a topic not substantively covered in existing literature, which is largely founded on data collected online, such as in the process of optimising advertising revenue. The adverts you are delivered might slowly change your behaviour, but each one can hardly be thought to have a significant impact. This is not the case in the public sector. As [X8] recalled above when discussing crime dispersion, feedback effects in practice can be so strong that they make models rapidly fail. The effect of this property on fairness and accountability in systems has yet to be properly unpacked and explored in context.

If the way that data changes in practice is of concern, how might we respond to it? One way to see this issue is from an interpersonal perspective. The idea of a *visibility debt* has been raised occasionally before and written about by engineers inside machine learning and in traditional software engineering.¹⁰²⁵ To the upstream data collectors, there are *undeclared users* of their data streams. To the downstream users, there are individuals exerting influence over their system that that might not even be

¹⁰²² Seaver, ‘Knowing Algorithms’ (n 52).

¹⁰²³ David Sculley and others, ‘Hidden technical debt in machine learning systems’ in *Proceedings of the 28th International Conference on Neural Information Processing Systems, Montréal, Canada – December 07 - 12, 2015* (Cambridge, MA, 2015).

¹⁰²⁴ See eg Ensign, Friedler, Neville, Scheidegger and Venkatasubramanian (n 48).

¹⁰²⁵ See J David Morgenthaler, Misha Gridnev, Raluca Sauciu and Sanjay Bhansali, ‘Searching for Build Debt: Experiences Managing Technical Debt at Google’ in *Proceedings of the Third International Workshop on Managing Technical Debt (MTD '12)* (IEEE Press 2012); Sculley and others (n 1023).

4. Coping with Value(s) in Public Sector Machine Learning

aware that such a system exists, particularly when data is collected in a decentralised manner by, say, auditors or police patrol officers. This problem is only going to be exacerbated as more models are made using the same upstream data sources, and bi-lateral communication becomes more and more challenging. Better communication might help, but must overcome difficult hurdles of explaining to upstream actors the kind of changes that matter downstream, and the kind that don't, in ways that they not only understand (as they might be relatively statistical) but that they can identify and act on within their roles. This is all compounded by how changing upstream data collection might not be an explicit act at all, but one emerging from cultural change or use of discretion. This is emphasised in the importance of so-called 'street-level ministers' in the public administration literature, which points out how formal rules are only part of the picture, and that many day-to-day choices in the grey zones are made by bureaucrats at the frontlines of public service.¹⁰²⁶ Where change does occur, managers might not notice it, as in their day-to-day roles or through their monitoring and evaluation tools, they only see part of the picture.¹⁰²⁷

A second way to look at the problem assumes communication failure is inevitable, so focuses on the changing data itself. This would necessarily involve concept drift detection, sets of techniques designed to automatically detect shifts in distributions that might be relevant to a modelling task. Concept drift detection, particularly in complex real-world contexts, is difficult and daunting theoretically, let alone practically.¹⁰²⁸ Some of the more recent reviews in the field call for the integration of domain knowledge in order to discern relevant drift,¹⁰²⁹ yet there are few, if any, well-explored methods for doing this.

4.7.2. 'Always a person involved': Augmenting outputs

While we hear horror stories of the results of algorithms unquestioningly replacing swathes of existing analytical practice and institutional knowledge, our informants' experiences do not reflect that. Many organisations interviewed here have well-developed routines for augmenting algorithmic outputs, such as crime maps, with contextual data using manual analysts. As one informant described above, their 'predictive policing' system was not supposed to bring in shocking new insights, but relieve analysts from the slog of generating maps so that they could get on with more

¹⁰²⁶ See eg Lipsky (n 961).

¹⁰²⁷ See Aurélien Buffat, 'Street-level bureaucracy and e-government' (2015) 17(1) *Public Management Review* 149 DOI: 10/gfgrbf.

¹⁰²⁸ See generally Quiñonero-Candela, Sugiyama, Schwaighofer and Lawrence (n 366); Gama, Žliobaitė, Bifet, Pechenizkiy and Bouchachia (n 366).

¹⁰²⁹ See Gama, Žliobaitė, Bifet, Pechenizkiy and Bouchachia (n 366).

advanced work. How algorithmic systems are examined day-to-day and how humans enter ‘the loop’ of decision-making at different stages is an important area for future design focus. There are many points for intervention in a decision support system outside of the modelling process—for example, in the training data (many systems attempting to make fairer machine learning system intervene at this point¹⁰³⁰) or after the model has been generated¹⁰³¹), such as the stage between model output and map dissemination.¹⁰³² Particularly in this latter stage, design interventions are likely to be key. If a statistical definition of fairness is reached, it may be possible to make a ‘fair’ model, for example by introducing fairness constraints to optimisation. This provides no guarantees about decision-support being interpreted fairly. Designers should not just consider how to design artifacts such as maps to promote fairness, but should also do so in contexts imagining that models have been ‘scrubbed’ of certain types of bias, to understand if this introduces any additional effects. In the messy outside world, these efforts may interact, and it is not guaranteed that the sum of two good efforts is also effective.

Taking this areas forward will likely require building upon and rethinking traditional knowledge elicitation techniques. Effective knowledge elicitation, as part of the hot topic of knowledge acquisition in heady days of expert systems, was thought to be a foundational building block of AI.¹⁰³³ With the inductive, data-driven turn, we may need to rediscover it as something which constrains and augments patterns learned from data, less around tough or rare cases¹⁰³⁴ as much as around contentious, value-laden ones. This will require very different sorts of prioritisation and elicitation methods than developed so far, and seems a promising and urgent avenue for future research.

4.7.3. ‘When it aligns with intuition’: Understanding discretion

It is commonly claimed that people over-rely on algorithmic systems, or increasingly consider them neutral or authoritative.¹⁰³⁵ I do not claim this is not an issue—but ac-

¹⁰³⁰ See eg Faisal Kamiran and Toon Calders, ‘Data preprocessing techniques for classification without discrimination’ (2012) 33(1) Knowledge and Information Systems 1 DOI: 10/b36t4t; Feldman, Friedler, Moeller, Scheidegger and Venkatasubramanian (n 237), in addition to section 1.6.2, p. 65.

¹⁰³¹ See eg Faisal Kamiran, Toon Calders and Mykola Pechenizkiy, ‘Discrimination aware decision tree learning’ in *2010 IEEE International Conference on Data Mining* (2010) DOI: 10/bqdjmp, in addition to section 1.6.2, p. 65.

¹⁰³² See eg Hsinchun Chen and others, ‘Visualization in law enforcement’ in *CHI’05 Extended Abstracts on Human Factors in Computing Systems* (ACM 2005).

¹⁰³³ See eg Cooke (n 168); Robert R Hoffman, ‘Human Factors Contributions to Knowledge Elicitation’ (2008) 50(3) Human Factors 481 DOI: 10/ccd4v3, as well as above in section 1.6.1.

¹⁰³⁴ See Hoffman, Crandall and Shadbolt (n 168).

¹⁰³⁵ See eg danah boyd, ‘Undoing the neutrality of Big Data’ (2016) 16 Florida Law Review Forum 226.

4. Coping with Value(s) in Public Sector Machine Learning

According to the informants in this project, this framing is one-dimensional. In particular, if and how individuals trust and rely on decision-support systems seems highly contextual in nature. The design strategies used to improve uptake of these systems, such as presenting prioritised lists or options, are understudied in relation to how these affect the mental models constructed by those using these systems day-to-day.

The way that different tasks, stakes or contexts mediate these effects is even less studied. We might expect there to be a difference in the perception of ‘neutrality’ of algorithms between those that direct police helicopters and those that flag children at risk of abuse; two very different tasks. We might not expect however, as informants reported, there to be a significant difference in the way algorithms were considered by helicopter pilots versus by police motorcyclists. Research in risk perception by helicopter pilots has found additional disparities between experienced and inexperienced users which is also worth unpacking in this context.¹⁰³⁶ Ultimately, to make blanket and somewhat alarmist statements about how algorithms are or are not being questioned is likely to alienate practitioners who recognise a much more nuanced picture on the ground, and hinder co-operation in this space between researchers and those who would benefit from research uptake.

As well as the demographics and contexts of when algorithms are trusted more or less on aggregate, we might be interested in patterns of over- or under-reliance *within* individuals or use settings. If users of decision-support choose to ignore or to follow advice at random, we may not be wholly concerned with this, or at least our concern might centre on the dimension of increasing adherence. Yet if there are *systematic* biases in the way that advice is or is not used—particularly if they result in individuals holding different protected characteristics being treated differently—then this may create cause for alarm, or at least merit further study. Assuming a system can be ‘scrubbed’ of bias and then forced onto users to obey is clearly not what will happen in real world deployments.

Lastly, researchers considering human factors in computer security have emphasised ‘shadow security practices’, which consist of ‘workarounds employees devise to ensure primary business goals are achieved’ and ‘reflect the working compromise staff find between security and “getting the job done”’.¹⁰³⁷ Similarly, studies of fairness and accountability in socio-technical systems must incorporate an assumption that there will be a mixture of technological resistance and ad-hoc efforts, which, similarly to the findings in human factors of security, will surely be ‘sometimes not as

¹⁰³⁶ Mary E Thomson, Dilek Önköl, Ali Avciođlu and Paul Goodwin, ‘Aviation risk perception: A comparison between experts and novices’ (2004) 24(6) Risk Analysis 1585 DOI: 10.1111/j.0272-4332.2004.00552.x.

¹⁰³⁷ See Iacovos Kirlappos, Simon Parkin and Angela Sasse, ‘“Shadow Security” As a Tool for the Learning Organization’ (2015) 45(1) SIGCAS Comput. Soc. 29 DOI: 10/gfgrbs.

secure as employees think.’ You can’t engineer ethics, and you can’t expect some individuals not to try, rigorously or not, to uphold it in ways they see fit. It is a useful heuristic to assume systems are trained on ‘pure’ streams of data and then must be cleaned of bias downstream, but in real data collection environments, even upstream actors in the data collection process attempt to work in the discretionary places computer systems allow (and create) to inject fairness where they see fit.¹⁰³⁸

4.7.4. ‘I’m called the single point of failure’: Moving practices

Most of the work in discrimination-aware data mining involves statistical assurance of fairer systems, or the installation of interfaces to make them more transparent. Most of the experiences of informants in this study were the opposite—social detection of challenges and social solutions to those challenges, none of which were mathematically demonstrated to work, but which organisationally at least were perceived to be somehow effective. Managing these challenges will require a balance between the two that has seldom been effectively struck. It seems unlikely that statistical practices could exist without the social practices, or the other way around.

This means that how the social practices are developed, maintained and transferred across contexts or over time is important to consider. Public sector bodies are under the constant shadow of their core quantitatively trained staff being poached, moving agencies, or leaving the sector entirely. Several interviewees had recently entered their job from another part of government where they pioneered analytics, or were about to leave from their current post. One modeller described how their manager called them ‘the single point of failure for the entire force’ [X15]. As discussed above, there is significant concern within the sector that less resourced sister organisations will import the models without the hard-won practices to understand and mitigate issues such as bias and discrimination. Some of the informal practices that are established might be able to be documented, at least for inspiration if not for reproduction—employee handover notes are of course commonplace in these organisations. Yet other practices, particularly any critical skills that led to the establishment of practices in the first place, will likely be more challenging to codify.

Encoding social practices that surround software systems has always been challenging. The stakes are now higher than ever. Part of these efforts might involve the creation of informal and dynamic knowledge-bases and virtual communities to share ethical issues and quandaries in relation to algorithmic support in practice,¹⁰³⁹ but

¹⁰³⁸ See Frans Jorna and Pieter Wagenaar, ‘The ‘Iron Cage’ Strengthened? Discretion and Digital Discipline’ (2007) 85(1) Public Admin. 189 DOI: 10/dqn2m2.

¹⁰³⁹ See Michael Veale and Reuben Binns, ‘Fairer machine learning in the real world: Mitigating discrimin-

these are unlikely to arise organically in all fields, particularly resource-scarce ones. Considering what collaborative work, co-operation and dissemination in these fields could and should look like is of immediate importance to practitioners today.

4.7.5. ‘Looks like we’ve 100% accuracy’: Talking performance

Some of the most value laden aspects of machine learned models relate to loss functions and performance metrics. Yet beyond accuracy, and perhaps false positives and false negatives, it becomes increasingly difficult to explain performance metrics effectively to individuals without technical backgrounds but that may either have vertical accountability for the project or necessary, extensive domain knowledge. As recalled above, some informants complained of challenges explaining performance when accuracy was not the appropriate task-specific metric, such as in heavily imbalanced datasets (where you can get a high accuracy by using a dumb classifier that always predicts one class). There are a range of performance metrics suitable for imbalanced data,¹⁰⁴⁰ but these mostly lack clear analogies for laypeople. Moving away from binary classification, explaining performance metrics for continuous regression tasks or multiple classification tasks is arguably more challenging still.

In other cases described above, the performance judged in other ways: models were not trusted or thought valuable if they did not contain individuals’ ‘preferred risk indicators’ [X2]; if they were too similar to analysis that existed before [X6]; or even if they were *more* accurate than was initially planned for, as the commissioners would rather the rest of that performance be substituted for interpretability [X3]. Other informants emphasised the importance of talking to the users before coming up with performance metrics [X16], as in some cases only actionable knowledge is worth optimising for.¹⁰⁴¹ This broadly chimed with many respondents’ conception of the most important performance metric of all—for contractors, whether a client bought a model, and for public servants or in-house modellers, whether their department actually used it.

Given that performance metrics are one of the most value-laden parts of the machine learning process,¹⁰⁴² it will be key to discuss them both with statistical rigour and with practical relevance. This intuitively seems to present domain-specific challenges in training, visualisation, user interfaces, statistics and metrics, problem struc-

ation without collecting sensitive data’ (2017) 4(2) *Big Data & Society* DOI: 10/gdcfnz.

¹⁰⁴⁰ See Japkowicz and Shah (n 231).

¹⁰⁴¹ See also Monsuru Adepeju, Gabriel Rosser and Tao Cheng, ‘Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions – a crime case study’ (2016) 30(11) *International Journal of Geographical Information Science* 2133 DOI: 10.1080/13658816.2016.1159684.

¹⁰⁴² Solon Barocas and Andrew D Selbst, ‘Big Data’s Disparate Impact’ (2016) 104 *California Law Review* 671 DOI: 10/gfgq9w; Cary Coglianese and David Lehr, ‘Regulating by Robot: Administrative Decision Making in the Machine-Learning Era’ (2016) 105 *Geo. LJ* 1147 (<https://ssrn.com/abstract=2928293>).

turing and knowledge elicitation, among other fields.

4.8. Interim conclusions

Researchers should be wary of assuming, as seems often the case in current discourse, that those involved in the procurement and deployment of these systems are necessarily naïve about challenges such as fairness and accountability in the public sector's use of algorithmic decision support. This assumption sits particularly uncomfortably with the value attributed to participatory design and action research in HCI and information systems.¹⁰⁴³ While those involved in acquiring these technologies for the public sector might not be prime candidates for developing new statistical technologies for understanding bias and outputs in complex models, this does not mean that they do not care or do not try to tackle ethical issues that they perceive. Indeed, as well as the individual perspectives in this paper, some public agencies are already developing their own in-house ethical codes for data science activities.¹⁰⁴⁴ Yet issues like fairness have been shown to come with technically difficult to reconcile, or even irreconcilable trade-offs—something well-demonstrated by the impossibility theorem illustrating that independently plausible formal definitions of fairness can be statistically incompatible with one another,¹⁰⁴⁵ or concerns raised that explanation facilities might work better for some outputs than for others.¹⁰⁴⁶ Reconciling these harder boundaries and issues within messy organisational contexts will present a major challenge to research uptake in this field in the coming years.

Where to go from here? The challenges outlined above—dealing with changing data, better understanding discretion and the augmentation of model outputs, better transmission of social practices and improved communication of nuanced aspects of performance—sit amongst a range of promising areas for future interdisciplinary collaboration. The implicit and explicit assumptions of proposed solutions to both these challenges and to the broader issues must be stress-tested in real situations. This presents important questions of methodology. Domain-specific, organisational and contextual factors are crucial to closely consider in the context of interventions intended to improve the fairness and accountability of algorithmic decision-support. The institutional constraints, high stakes and crossed lines of accountability in the pub-

¹⁰⁴³ Gillian R Hayes, 'The relationship of action research to human-computer interaction' (2011) 18(3) ACM Transactions on Computer-Human Interaction (TOCHI) 15 DOI: 10.1145/1993060.1993065; Richard L Baskerville and A Trevor Wood-Harper, 'A critical perspective on action research as a method for information systems research' (1996) 11(3) Journal of Information Technology 235 DOI: 10/b3r58v.

¹⁰⁴⁴ Department for Digital, Culture, Media & Sport, *Data Ethics Framework* (n 16).

¹⁰⁴⁵ Chouldechova (n 86).

¹⁰⁴⁶ Edwards and Veale, 'Slave to the Algorithm?' (n 79) 60.

4. Coping with Value(s) in Public Sector Machine Learning

lic sector arguably presents even more reason to do so. Only so much can be learned from studying systems *in vitro*, even with access to impressive quantities of relevant, quality data with which to experiment. Those interested in transformative impact in the area of fair and accountable machine learning must move towards studying these processes *in vivo*, in the messy, socio-technical contexts in which they inevitably exist. Interventions will have to cope with institutional factors, political winds, technical lock-in and ancient, withering infrastructure head on, as they would have to in the real world. Researchers will have to facilitate the navigation of contested values, and will not always have the freedom of seeking the types of accountability or fairness that they feel most comfortable with. Such challenges should be embraced. To enable this, trust will need to be built between public bodies and researchers; trust that is currently being endangered by 'gotcha!'-style research that seeks to identify problematic aspects of algorithmic systems from afar without working collaboratively to understand the processes by which they came about and might be practically remedied. Action research is a core methodology that would support these aims,¹⁰⁴⁷ but the combination of high stakes and a wariness that researchers might be spending more effort looking for algorithmic harms than offering help to fix it might make public agencies reluctant to open up to research interventions.

I believe that many of the individual issues highlighted above by the informants in this study could be developed into a novel strand of computer science or HCI inquiry. In the next section, I expand one of them to illustrate some of the challenges that arise when that is done.

¹⁰⁴⁷ Baskerville and Wood-Harper (n 1043).

5. Unpacking a tension: ‘Debiasing’, privately

Universal panaceas to value-laden, sociotechnical problems should be treated with suspicion. We are likely stuck with layered, messy techniques to define, resolve and manage these complex challenges. This section zooms in to examine one piece of this challenge—how potentially unfair patterns in datasets that make their way into modelling and decision-making processes might be remedied in practical rather than theoretical machine learning situations. I emphasise the situations already discussed in earlier sections¹⁰⁴⁸ where actors designing and deploying such systems wish to avoid bias themselves, for regulatory and reputation-related reasons, rather than adversarial situations (often within business models) where external investigators wish to discover bias against the will of the organisations undertaking analysis.¹⁰⁴⁹ I am further motivated by insight from the previous chapter, where alternative, external sources of information were drawn upon to try and understand patterns of potential discrimination, rather than examining the issues ‘within’ the dataset using the variables it contains.¹⁰⁵⁰

5.1. Knowing protected characteristics: necessary but problematic

To see why knowledge of protected characteristics is necessary, it is helpful to consider why certain naïve approaches to removing bias from modelling are inadequate. One could simply delete any sensitive variables related to discrimination, eg age, gender, race, or religion, from the training data. Unfortunately, this does not guarantee non-discrimination in the models that are trained on this data, as non-discriminatory items might exist which in some conditions are closely correlated with the sensitive

¹⁰⁴⁸ In particular, the interview findings around public sector machine learning in section 4.

¹⁰⁴⁹ This is not to say that the public sector is free from cover-ups, but that, as the public administration literature has emphasised, in many cases aims can be viewed within the public sector values frameworks in section 4.3.

¹⁰⁵⁰ See quote and finding on page 244.

5. Unpacking a tension: ‘Debiasing’, privately

attributes. Where geography serves as a sensitive proxy, this phenomenon is termed ‘redlining’. More broadly, it can be seen as an issue of redundant encoding.

In order to discover redlining in training data, one needs to be able to find out whether sensitive attributes might be encoded by other, apparently benign ones. For instance, to discover whether ZIP codes in a dataset are correlated with eg race, it will be necessary to either have race as an attribute in the dataset, or to have background knowledge about the demographics of the areas in question (for instance, from census records). Proposed approaches to non-discriminatory machine learning assume that whoever is implementing the technique has access to the sensitive attributes which might be encoded.¹⁰⁵¹ Such access is necessary for assurance of computationally non-discriminatory models.¹⁰⁵²

Despite this, in many cases organisations deploying machine learning will lack this necessary access, often for legitimate reasons.

First, the collection of personal data inevitably creates privacy risks. Many organisations have internalised the dictum of regulators and privacy advocates only to collect data that is necessary for their purposes. The concepts of data minimisation and purpose limitation within the GDPR are intended to prevent collection and processing of data for unspecified or disproportionate ends. Furthermore, the kinds of protected characteristics involved in cases of discrimination raise higher privacy and data protection risks than other kinds of data, and are given special protection under the GDPR, as described earlier.¹⁰⁵³ The proposition that organisations ought to collect a wide range of sensitive data that isn’t directly necessary for their primary purposes contradicts this general dictum. Yet fairness-aware machine learning seems to require organisations to do exactly that to adequately inspect and modify their models.¹⁰⁵⁴

It is not the aim here to analyse the extent to which privacy and data protection law and best practice is substantively in conflict with the collection and processing of sensitive attributes for the purposes of fairness-aware machine learning.¹⁰⁵⁵ Such work, while in scope of this thesis, will be left for another time or author. It may be that col-

¹⁰⁵¹ See eg Hajian and Domingo-Ferrer, ‘A Methodology for Direct and Indirect Discrimination Prevention in Data Mining’ (n 235); Hardt, Price and Srebro (n 241).

¹⁰⁵² Indrè Žliobaitė and Bart Custers, ‘Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models’ (2016) 24(2) *Artif. Intel. & Law* 183 DOI: 10/gfgt9b.

¹⁰⁵³ GDPR, art 9(1).

¹⁰⁵⁴ Some have argued that a principle of data minimumisation would enable better governance of these issues, rather than minimalisation. See Bart van der Sloot, ‘From Data Minimization to Data Minimumization’ in Bart Custers, Toon Calders, Bart Schermer and Tal Zarsky (eds), *Discrimination and privacy in the information society* (Springer 2012) DOI: 10/cwqq. In some ways, it could be argued that the existing regulation could be already read through such a lens, but text interpretation is not the focus of this section.

¹⁰⁵⁵ For some consideration relevant to that question, see Indrė Žliobaitė, Faisal Kamiran and Toon Calders, ‘Handling Conditional Discrimination’ in *2011 IEEE 11th International Conference on Data Mining* (IEEE 2011) DOI: 10/fzxwfp; Sloot (n 1054).

lection and processing for such purposes is legitimate;¹⁰⁵⁶ however, it may still not be desirable. It would require data subjects to share sensitive attributes along with non-sensitive ones every time their data was to be used to train a model. The general result would be much more sensitive data in the hands of data controllers—a security risk even if it is intended to be used for the legitimate purposes of avoiding discriminatory outcomes. Even if organisations are permitted to collect and process such data, requiring consumers to provide it might make their service less competitive, or less trusted. For purposes of building a model that serves some narrowly prescribed goal, they may not see the need to collect sensitive data. In the context of data minimisation, the data controller must argue that it is proportionate to collect and process sensitive categories of data, and they may not be sufficiently incentivised to do so. Where individuals fear they are being treated unfairly, the collection of sensitive data by the organisation in question, even to explicitly remedy fairness issues, might not alleviate that perception-based fear. It could even make it worse.

Some approaches have been proposed to transform training data with anonymisation procedures to protect the sensitive attributes. This can be performed in tandem with preprocessing techniques to prevent discrimination.¹⁰⁵⁷ While promising, this still mandates the comprehensive collection of sensitive attributes from individuals in training data for each form of discrimination for which mitigation is desired. Despite meaningful privacy protections, the concerns raised above are still likely to apply. Individuals are unlikely to be happy providing a comprehensive range of sensitive personal data to the very organisations who are in position to discriminate, no matter how technically robust their anonymisation process is.

5.2. ‘Debiasing’ with limited sensitive data

Organisations developing learning systems need strategies to mitigate discrimination concerns in the absence of sensitive data. The challenge is to implement the techniques, such as those outlined above, without having to take on the additional burden and risk of collecting detailed sensitive data on the training sample.

Here, I present three proposals to overcome this challenge. The first is based on a multi-party data governance model, suited to contexts where little background know-

¹⁰⁵⁶ Under, for example, an exemption in national law. One place to look might be the *Equality of opportunity or treatment exemptions* in DPA 2018, sch 1 para 8, although several provisions within that paragraph cloud its use for legitimising fair machine learning.

¹⁰⁵⁷ Hajian and Domingo-Ferrer, ‘A Methodology for Direct and Indirect Discrimination Prevention in Data Mining’ (n 235); Sara Hajian, Josep Domingo-Ferrer and Oriol Farràs, ‘Generalization-based privacy preservation and discrimination prevention in data publishing and mining’ (2014) 28(5-6) *Data Min. Knowl. Discov.* 1158 DOI: 10/f6fs97.

5. Unpacking a tension: ‘Debiasing’, privately

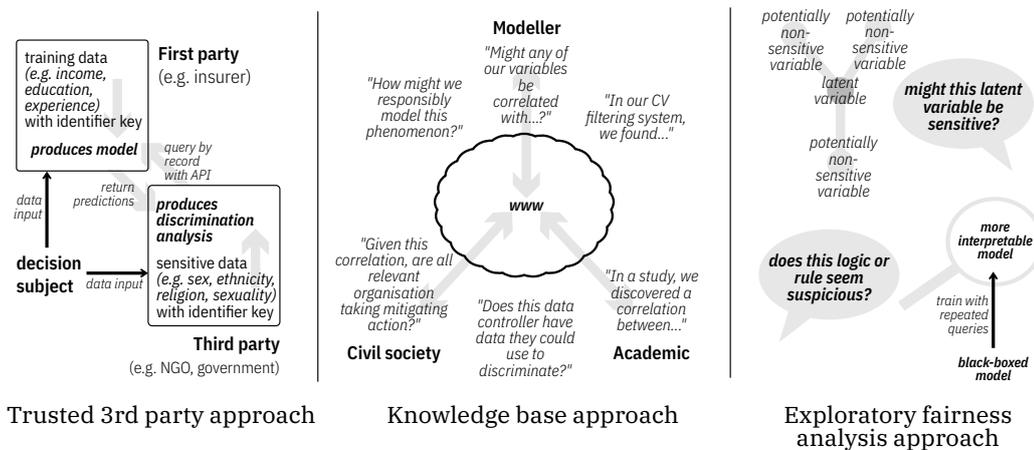


Figure 5.1.: Three approaches to ‘debiasing’ without holding sensitive characteristics.

ledge about discrimination exists and a comprehensive assessment of potential forms of discrimination is needed. The second is a collaborative knowledge sharing approach in which organisations can learn from each other’s experiences in similar contexts, as well as identify relevant sociological and demographic correlations. The third involves exploratory analysis to build hypotheses of potential unfair characteristics of the data or system, which can be more formally tested as part of a due diligence process. Figure 5.1 pictographically illustrates these three distinct approaches.

None of these three methods are perfect, nor do they provide complete solutions or assurances to the multitude of challenges surrounding machine learning systems. It is argued instead that these are avenues that are important to explore to make fairer machine learning a practical reality in the multitude of settings that automated and semi-automated decisions will be occurring in our society in the coming years and decades.

5.2.1. Trusted third parties

Various proposals have been made for the involvement of external parties in the evaluation and auditing of algorithmic systems.¹⁰⁵⁸ Some of these are reflected in law. As

¹⁰⁵⁸ See eg Alessandro Mantelero, ‘Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection’ (2016) 32(2) Computer Law & Security Review 238; Frank Pasquale, ‘Beyond innovation and competition: The need for qualified transparency in internet intermediaries’ (2010) 104(1) Nw. U. L. Rev. Andrew Tutt, ‘An FDA for Algorithms’ (2017) 69 Administrative Law Review 83; Sandvig, Hamilton, Karahalios and Langbort (n 506)

discussed previously,¹⁰⁵⁹ the GDPR obliges organisations to undertake data protection impact assessments (DPIAs) wherever ‘profiling’ is used to automatically make decisions which have legal or significant effects on data subjects. In some cases these assessments may be audited by a data protection authority.¹⁰⁶⁰ It is not uncommon in governance approaches involving audits for external auditors to be given access to an organisation’s policies, personnel, data collection procedures, training data, models, proprietary code, and other relevant aspects, in order to assess the ethical dimensions and legal compliance of the subject matter—in this case, a particular algorithmic system.

This model assumes that the relevant information required to perform an audit will lie in the hands of the organisation being audited. As argued above, this might not be the case, rendering external audit process incapable of ensuring the kinds of algorithmic fairness that DADM and FATML techniques aim for.

This might be different, were trusted third parties enlisted to work alongside organisations from when data collection begins. This proposal could be achieved with a variety of different institutional and technical arrangements. Below, I illustrate several possible implementations.

The first party (the organisation implementing the algorithmic decision-making system) has access to historical data relevant to the classification or prediction task for which they are building a model. However, the first party does not and *should* not have access to any of the protected characteristics associated with the population used to train the model.

As discussed above, in order to statistically test the model for potential discrimination, the protected characteristics need to be linked somehow to the records used in the training data. To achieve this, a trusted third party is enlisted to collect data on the protected characteristics of those individuals whose data is used to train the model. For each individual, protected characteristics like race, gender, religious beliefs or health status are collected by the third party in parallel to the collection of the non-protected characteristics by the third party. The channel for communicating this information from the individual to the third party may depend on the platform (eg online, telephone, or in-person). It could be as part of a separate collection process, although this prove unwieldy, or be encrypted simultaneously and seamlessly at the point of collection (eg locally through JavaScript web cryptography) with the public key of a third party, and transmitted to the organisation in question.¹⁰⁶¹

¹⁰⁵⁹ See section 2.4.2.

¹⁰⁶⁰ GDPR, recital 84.

¹⁰⁶¹ A range of privacy preserving communication solutions could be applicable here if the modeller was treated as an adversary. While the methods here implicitly focus on organisations actively wishing

5. Unpacking a tension: ‘Debiasing’, privately

Consider the following illustrative example:

An insurer wishes to use a machine learning model to help determine customers’ premiums. They have access to historical customer data, and use it to train a model to predict the amount of compensation a customer will claim over the term of their cover given certain attributes (eg postcode, occupation, qualifications). The estimated size of a potential claim—the output of the model—is used to automatically set premiums.

The insurer enlists a third party organisation (for instance, a consumer rights group) to simultaneously collect protected characteristics about each customer as they purchase their insurance policy. For online purchases, the customer is directed to the consumer rights group’s domain, and asked to provide protected characteristics for the purposes of discrimination prevention.

Based on this multi-party data governance model, there are multiple ways to proceed, depending on whether the goal is merely to detect bias or to both detect and prevent it, and what prevention techniques will be used (eg pre-processing, in-processing, or post-processing). I now outline a set of possible variations within this approach, and discuss their relative advantages and drawbacks. The first consists of the third party *detecting* discriminatory effect, while the second consists of the third party aiding in its *mitigation*.

5.2.1.1. As ex post discrimination detector

In cases where the third party’s only role is to detect discrimination (but not prevent it), the third party need only collect protected characteristics from each individual featured in the dataset used to train (and test) the model, along with an identifier. The records held by the first party for the purposes of model training could be linked by this identifier to the records held by the third party which contain the protected characteristics. The third party would be given access to the model developed by the first party (either directly or via an API). By testing the outputs of the model on each of the individuals in their sensitive attribute dataset (using the individual’s identifier), the third party could detect disparate impacts.

An advantage of this variation is that the third party can only access the sensitive attributes, not the potentially non-sensitive ones. Since each record only contains sensitive attributes and an identifier this represents a lesser privacy risk; while the

to increase trust and reduce discriminatory outcomes, a proposal the author worked on as part of a research team will be outlined below in section 5.2.1.4.

data itself is sensitive, it would be harder to re-identify an individual without other data types. This may also be beneficial from the perspective of a first party concerned about keeping their proprietary model secret, as it has been shown that unlimited access to a query interface for a prediction model can allow an attacker to extract and reconstruct a model.¹⁰⁶² In this case, while the third party would have unrestricted ability to query the model by individual identifiers, and thus learn the distributions of outputs for each protected characteristic, they would not be able to reverse-engineer the model without access to the other, non-protected characteristics.¹⁰⁶³

The disadvantage of this variation is that it only provides the first party with evidence of the *disparate impact* of their model. Disparate impact is a blunt measure of discrimination, because some disparities may be ‘explicable’, in the sense that the disparities might be accountable by reference to attributes which are legitimate grounds for differential treatment.¹⁰⁶⁴ Furthermore, measures of disparate impact may not be sufficient for the first party to actually change their model to prevent it from being discriminatory. For instance, to remove bias from the training data, the first party would have to know which data points to relabel, massage or re-weight—ie the protected characteristics of the specific individuals, which they would lack. More generally, without the ability to check for redundant encoding of protected characteristics by non-protected attributes, it will be difficult for the first party to revise their model.

Nevertheless, the mere ability to detect disparate impact may be valuable in allowing third parties to flag up problems, which can then be dealt with by allowing the first party access to the necessary additional data to investigate and transform their model accordingly. Separating out detection of disparate impact and prevention could thus prevent unnecessary sharing of sensitive attributes and enable the third party to perform continuous monitoring.

5.2.1.2. As ex ante discrimination mitigator

Alternatively, the third party could collect both the protected attributes and the other features used to train the model. This would enable the third party to play a more significant role, not only detecting disparate impact in model outputs but also helping to ensure the disparities are attributable to disparate mistreatment (ie that they are not explainable), and also to ensure that the model can be bias-free.

¹⁰⁶² Tramèr, Zhang, Juels, Reiter and Ristenpart (n 746).

¹⁰⁶³ For further discussion of confidentiality breaches involving machine learning systems, refer to section 3.2.

¹⁰⁶⁴ Zafar, Valera, Gomez Rodriguez and Gummadi (n 87); Žliobaitė and Custers (n 1052).

As redlining detector In this approach, the third party has both the sensitive and potentially non-sensitive characteristics, and puts them through a common framework to produce summary information that aims to flag obvious issues that might occur during model building. Upon acquisition of a cleaned dataset, the third party calculates and returns a set of redundant encodings and their strengths. The returning document might note that ‘race is correlated to zip code by 0.8’; ‘gender is correlated with aspects of profession by 0.2’, and so on. The first party could use this knowledge to make trade-offs in the model—removing certain features, or engaging in further discussions with the third party about potential procedures to scrub unwanted correlations from a model.

Naturally, such a framework could suffer from flaws which made it unsuitable for some types of data or problems, particularly highly contextual ones. Yet this approach would create a focal point for the improvement of discrimination detection methods for certain contexts and data types, which would foster active discussion and debate about best practices and processes that could be translated into on-the-ground practice with relative ease.

As data preprocessor Another approach would see the third party pre-process the training data in such a way as to preserve anonymity and remove bias, before handing it over to the first party. This could be achieved by modifying the data to preserve degrees of anonymity, using techniques such as statistical disclosure control¹⁰⁶⁵ and privacy-preserving data mining¹⁰⁶⁶ which allow the statistical properties of the data to be maintained, followed by applying one of a range of anti-biasing techniques described in the DADM/FATML literatures.¹⁰⁶⁷ It would even be possible, if it were desired, to introduce positive discrimination at this point, and some methods have been proposed for how this could be achieved.¹⁰⁶⁸ As mentioned above, more recently proposed techniques aim to render datasets both k -anonymous and non-discriminatory in a single procedure with limited loss of accuracy.¹⁰⁶⁹ Having transformed the data

¹⁰⁶⁵ Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer and Peter-Paul de Wolf, *Statistical Disclosure Control* (John Wiley & Sons 2012); Leon Willenborg and Ton de Waal, *Elements of Statistical Disclosure Control* (Springer 2012).

¹⁰⁶⁶ Rakesh Agrawal and Ramakrishnan Srikant, ‘Privacy-preserving Data Mining’ in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (ACM 2000).

¹⁰⁶⁷ See eg Feldman, Friedler, Moeller, Scheidegger and Venkatasubramanian (n 237); Hajian and Domingo-Ferrer, ‘A Methodology for Direct and Indirect Discrimination Prevention in Data Mining’ (n 235); Kamiran and Calders (n 1030).

¹⁰⁶⁸ Sicco Verwer and Toon Calders, ‘Introducing positive discrimination in predictive models’ in Bart Custers, Toon Calders, Bart Schermer and Tal Zarsky (eds), *Discrimination and privacy in the information society* (Springer 2013).

¹⁰⁶⁹ Hajian, Domingo-Ferrer and Farràs (n 1057); Sara Hajian and Josep Domingo-Ferrer, ‘Direct and indirect discrimination prevention methods’ in Bart Custers, Toon Calders, Bart Schermer and Tal Zarsky (eds), *Discrimination and privacy in the information society* (Springer 2012).

to increase privacy and remove bias, the third party could then hand it over to the first party for model development.

The advantage of this variation is that the first party can develop whatever kind of model they like, without the risk of it learning biases from the training data. It also limits the involvement of the third party to a single step, after which the data could be deleted. Finally, it encourages the development of expertise on the part of the specialist third party and doesn’t require the first party to have in-house knowledge about fairness-aware machine learning. The disadvantage of this approach is that the anonymisation techniques only provide a degree of (quantifiable) anonymity. There is a clear trade-off between degrees of anonymity and utility of the dataset,¹⁰⁷⁰ such that useful datasets will still likely carry re-identification risks. To the extent that such risks persist, the first party could learn more about individuals’ sensitive characteristics in this variation than it could in the other variations.

5.2.1.3. Who could be a third party?

I have thus far assumed the existence of a suitable trusted third party, but it is worth considering what kinds of organisations might fulfil this role. This will likely depend on which of the variations are adopted. Each might pose different requirements of trustworthiness, technical expertise and incentivisation. In the case of a third party whose role is merely to detect disparate impact, relatively little technical expertise would be required, making it suitable for organisations with fewer resources and technical skills. The fact that disparate impact is already the focus of many civil society groups’ research activities may make them well situated to take on this role. Many potentially affected minority groups already have active representatives who could benefit from more formal auditing roles. Depending on the application context, it may be appropriate to involve different organisations; for instance, trade unions might be more equipped to address the fairness of algorithmic models deployed in human resources decisions.

If the third party is expected to be an ex ante discrimination mitigator, they will require more data collection and particular expertise in fairness-aware techniques. It may therefore need to be a specialist organisation, potentially working in collaboration with appropriate civil society organisations. It could be anticipated that consultancy or accountancy firms might provide these services to corporate clients, as they do with other forms of social auditing.

¹⁰⁷⁰ Grigorios Loukides and Jianhua Shao, ‘Data utility and privacy protection trade-off in k-anonymisation’ in *Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society* (ACM 2008).

5. Unpacking a tension: ‘Debiasing’, privately

New bodies and actors designed to take a cross-cutting role in data processing and analysis are emerging. Following a report from the Royal Society and British Academy suggesting a ‘data stewardship body’,¹⁰⁷¹ the UK Government has established a ‘Centre for Data Ethics and Innovation’ as an arms-length body (a ‘quango’) from the Department for Digital, Culture, Media and Sport (DCMS). The exact terms of this body are yet to be released at the time of writing but it is proposed that this centre ‘support the government to enable safe and ethical innovation in the use of data and AI’ through i) identifying steps to ensure that the law, regulation and guidance keep pace with developments in data-driven and AI-based technologies; ii) publishing recommendations to government on how it can support safe and ethical innovation in data and AI; and iii) providing expert advice and support to regulators (including for example the ICO, Competition and Markets Authority (CMA) and sectoral regulators) on the implications of data and AI uses and areas of potential harm. It is also proposed that such a body have a statutory footing.¹⁰⁷² Bodies with comparable cross-cutting competencies can be found emerging elsewhere, such as the French National Digital Council (*Conseil national du numérique*) and the recent German Data Ethics Commission (*Datenethikkommission*), designed to ‘develop ethical guidelines for the protection of individuals, the preservation of the structure of social life and the safeguarding of prosperity in the information age’.¹⁰⁷³

Another option might be statutory or chartered bodies whose remit includes monitoring discrimination, promoting equality, or enforcing law. For instance, the Equality and Human Rights Commission in the UK, or the Equal Employment Opportunity Commission in the US, are statutory bodies responsible for enforcing equalities laws. While traditionally involved in reviewing of individual cases for litigation, providing legal assistance and intervening in proceedings, these bodies could also take on more ongoing, data-driven monitoring of data-driven discrimination. State-sponsored API frameworks such as GOV.UK Verify,¹⁰⁷⁴ where the public sector certifies companies to provide verification services to third parties, might also serve as a framework to allow auditors to query trusted bodies for protected characteristics.

¹⁰⁷¹ The Royal Society and the British Academy, *Data management and use: Governance in the 21st Century* (The Royal Society and the British Academy 2017). The author of this thesis was drafting author of this report.

¹⁰⁷² Department for Digital, Culture, Media & Sport, *Centre for Data Ethics and Innovation Consultation* (HM Government 2018) (<https://perma.cc/GG22-GA4Q>) accessed 6th October 2018.

¹⁰⁷³ Bundesministerium des Innern Für Bau und Heimat, *Datenethikkommission* (Government of Germany 2018) (<https://perma.cc/7RC6-J8SD>) accessed 6th October 2018.

¹⁰⁷⁴ Although state support for this has recently been withdrawn.

5.2.1.4. Cryptographic ‘third parties’

Another option is possible, and one which this author and colleagues have proposed in other work not integrated into this thesis.¹⁰⁷⁵ It is possible, as we outline in that work, to cryptographically undertake auditing or ‘debiasing’ methods using encrypted sensitive characteristics. In the set-up we outline, two or more actors which include the organisation that wishes to produce a ‘debiased’ model engaged in a form of collaborative computation called *secure multiparty computation*. This technique effectively allows a computation to be completed despite the data used not being available to any one actor in cleartext. In the model we propose there, users hand over their ‘non-sensitive’ data to the modeller as in the set-ups assumed above, but hand over their sensitive category such a race or health data in an encrypted form, effectively splitting it between two or more actors, one of which could be a regulator with a statutory obligation not to collude with the modellers. Through secure multiparty computation, this data can be included in model training or auditing methods without any of the actors being able to view it. Secure multiparty computation is computationally intensive and can be difficult to co-ordinate however, and this method will likely need further technical development before it is deployable in practice.

5.2.2. Fairness knowledge bases

Experiential knowledge concerning the construction or attempted construction of ethical algorithmic systems has been largely neglected in the DADM and FATML communities. This has created a knowledge gap—not an insignificant one—that I believe has problematic consequences on-the-ground. Part of this gap has already been analysed above in chapter 4. This neglect is surprising for several reasons.

As data governance tools move increasingly towards ex ante prevention and anticipation of harms, particularly through data protection and privacy impact assessments,¹⁰⁷⁶ relying solely on in-data analysis of unfairness appears not just at tension with on-the-ground regulatory needs—it could even be described as paradoxical. It certainly seems problematic to have to link the data and train a system before you can decide whether you should even be doing either of those things. Many organisations cannot legally or practically proceed with any data work, even basic data access, cleaning, linking or exploration, until this stage is passed. Yet DADM and FATML ap-

¹⁰⁷⁵ Kilbertus, Gascon, Kusner, Veale, Gummadi and Weller (n 263).

¹⁰⁷⁶ 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); Reuben Binns, Michael Veale, Max Van Kleek and Nigel Shadbolt, ‘Like trainer, like bot? Inheritance of bias in algorithmic content moderation’ in Giovanni Luca Ciampaglia, Afra Mashhadi and Taha Yasseri (eds), *Social Informatics: 9th International Conference, SocInfo 2017, Proceedings, Part II* (Springer 2017) DOI: 10/cvc2.

5. Unpacking a tension: ‘Debiasing’, privately

proaches often implicitly assume that all the ingredients are on the table to build the tool, and the only decision to be made is whether to deploy or not.

Machine learning is a generic technology with sector-specific applications. High profile, consequential domains have included anticipating the geospatial distribution of crime,¹⁰⁷⁷ the need for child protection,¹⁰⁷⁸ and the detection of tax fraud.¹⁰⁷⁹ Some ethical issues are sector- or even location-specific, but others are likely to be shared. Highly problematic issues might only appear rarely, limiting their propensity to capture with in-data analysis.

Limited implementation and education surrounding DADM and FATML technologies threatens our ability to cope with pressing issues in today’s machine learning systems. Even though this research field has some history,¹⁰⁸⁰ usable software libraries remain largely unavailable, and little training for them exists.¹⁰⁸¹ Given the current lack of practical ethics education in computer science curricula,¹⁰⁸² rapid change seems unlikely. A stopgap is sorely needed.

Diagnosing and addressing social and ethical issues in machine learning systems can be a high capacity task, and one difficult to plan and execute alone or from scratch. Ethical challenges or appropriate methods to tackle them might lurk within aspects of processing that are easy overlooked, such as hyperparameters, model structure, or quirks in data formatting or cleaning. Some issues that might arise might also not have their origins in the models or the data, but surrounding social, cultural and institutional contexts. Issues such as automation bias,¹⁰⁸³ where individuals either place too

¹⁰⁷⁷ Azavea, *HunchLab: Under the Hood* (Author 2015) (<http://cdn.azavea.com/pdfs/hunchlab/HunchLab-Under-the-Hood.pdf>); Walter L Perry, Brian McInnis, Carter C Price, Susan C Smith and John S Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (RAND Corporation 2013) (http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf).

¹⁰⁷⁸ Cuccaro-Alamin, Foust, Vaithianathan and Putnam-Hornstein (n 981); Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko and Rhema Vaithianathan, ‘A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions’ in *Conference on Fairness, Accountability and Transparency (FAT* 2018)* (2018).

¹⁰⁷⁹ ‘Building and integrating databases for risk profiles in the United Kingdom’, in M S Khwaja, R Awasthi and J Loeprick (eds), *Risk-based tax audits: Approaches and country experiences* (World Bank 2011); Anuj Sharma and Prabin Kumar Panigrahi, ‘A review of financial accounting fraud detection based on data mining techniques’ (2012) 39(1) *IJCAI* 37.

¹⁰⁸⁰ See older work including Andrews, Diederich and Tickle (n 180); Dino Pedreschi, Salvatore Ruggieri and Franco Turini, ‘Discrimination-aware data mining’ in *ACM KDD ’08* (ACM 2008) DOI: 10/c7xx96.

¹⁰⁸¹ One interesting initiative to try and increase the ease of getting into the field is the draft textbook: Solon Barocas, Moritz Hardt and Arvind Narayanan, *Fairness and Machine Learning* (fairmlbookorg 2018) (<http://www.fairmlbook.org>).

¹⁰⁸² Michael Goldweber, Renzo Davoli, Joyce Currie Little, Charles Riedesel, Henry Walker, Gerry Cross and Brian R Von Kinsky, ‘Enhancing the social issues components in our computing curriculum: Computing for the social good’ (2011) 2(1) *ACM Inroads* 64; Carol Spradling, Leen-Kiat Soh and Charles Anson, ‘Ethics training and decision-making: do computer science programs need help?’ in *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education* (ACM 2008) vol 40.

¹⁰⁸³ Skitka, Mosier and Burdick (n 232).

much trust or too little trust in decision support systems, might be a synergistic result of both the model and the user interface. Evidence in previous chapters of this thesis points to that.¹⁰⁸⁴ Other issues might have their origins in a model but likely solutions elsewhere. For example, for fairness grievances which are particularly difficult to detect or anticipate, better systems for decision subjects to feedback to decision-makers might be required. These issues might not have one-size-fits-all answers, but they are also unlikely to need to be treated as fresh each and every time they arise.

Issues of changing data populations and correlations are both currently under-emphasised in DADM/FATML work and appear difficult to fully address with in-data analysis. This has similarly been highlighted in the interview work earlier in this thesis.¹⁰⁸⁵ Concept drift or dataset shift refers to either real or virtual (differently sampled) changes in the conditional distributions of model inputs and outputs¹⁰⁸⁶—for example, how changes in law might qualitatively affect prison population or the strategies of fraudsters. Fairness and transparency are not static but moving targets, and ensuring their reliability is important. But anticipating change is technically difficult. Knowledge around rates and causes of change can be tacit, obliging us to carefully consider how best to use expert input.¹⁰⁸⁷ In particular, these phenomena can be hard to examine when changes are nuanced, or even are a result of the actions of previous machine learning supported decisions themselves. An important key role for domain experts going forward would be to explain and record how and why certain types of concept drift occur, rather than just help in their detection.

5.2.2.1. Practical considerations

Given the above factors, I propose that a structured, community-driven data resource containing practical experiences of fair machine learning and modelling could serve as a useful resource both in the direct absence of sensitive data, and more broadly in its own right. Such a resource, likely held online, would allow modellers to record experiences with problematic correlations and redundant encoding while modelling certain phenomena, as well as sociotechnical ethical issues more broadly (such as interpretability, reliability and automation bias), and detail the kinds of solutions and approaches they used or sought to remedy them. It could operate on a relatively open, trust-based model, such as Wikipedia, or have third-party gatekeepers, such as NGOs or sectoral regulators verifying contributions and attempting to instil anonymity

¹⁰⁸⁴ See section 4.6.1.2.

¹⁰⁸⁵ See section 4.6.2.4.

¹⁰⁸⁶ Quiñonero-Candela, Sugiyama, Schwaighofer and Lawrence (n 366).

¹⁰⁸⁷ Gama, Žliobaitė, Bifet, Pechenizkiy and Bouchachia (n 366).

5. Unpacking a tension: ‘Debiasing’, privately

where possible or desired. It would create a stepping-stone to enable practical, albeit rudimentary, fairness evaluations to be carried out today.

Linked data technologies have already seen significant adoption in sectors where cross-organisational collaboration around data is necessary.¹⁰⁸⁸ This does not necessarily mean an industry-wide, comprehensive, rigid ontology for the purposes of addressing the ethical challenges of machine learning has to be adopted. Rather, a minimal adoption of common practices would enable different organisations to collaboratively annotate and describe the resource.

Several challenges would need to be addressed before such a database could be implemented. Similar variables and entities would need to be aligned in order to make such a dataset structured and navigable. Higher level common identifiers might be needed to group variables even if the levels of such variables were different. Some categorisations might have given individuals the chance to specify non-binary gender identities, or to opt out from this question—but this is unlikely to make any correlations or lessons found completely irrelevant or non-transferable in practice. Database ontologies should incorporate broader parts of the modelling process, such as cleaning or user interfaces, but the best format to do this is unclear. Arriving at it will likely be a result of trial-and-error.

Metadata should also be standardised. What kind of discrimination discovery methods were being utilised? How could effect strength or statistical significance be captured across these? It is likely that a descriptive vignette would also be useful, particularly concerning social processes and organisational context, but should or could this take a standardised format whilst remaining effective?

Such a dataset might benefit from discussion and input from different viewpoints both within the organisations submitting the information, but also externally. Open annotation or discussion technologies might contribute questions and context to the methods and content of dataset entries.¹⁰⁸⁹ Platforms such as *StackExchange*, a question and answer network initially aimed at developers, but recently with wider adoption, have proved practically popular technical and social tools for solving issues around software. Such a database could take inspiration from the factors that make knowledge communities run effectively in these virtual environments. Allowing organisations to trace the sources of the data in such collaborative knowledge bases

¹⁰⁸⁸ Christian Bizer, Tom Heath and Tim Berners-Lee, ‘Linked data—the story so far’ in Amit Sheth (ed), *Semantic services, interoperability and web applications: Emerging concepts* (IGI Global 2009).

¹⁰⁸⁹ See eg Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner and Lydia Pintscher, ‘From Freebase to Wikidata: The Great Migration’ in *Proceedings of the 25th International Conference on World Wide Web (WWW 2016)* (International World Wide Web Conferences Steering Committee 2016); Elena Simperl and Markus Luczak-Rösch, ‘Collaborative ontology engineering: A survey’ (2014) 29(1) *Knowl. Engin. Rev.* 101; Denny Vrandečić and Markus Krötzsch, ‘Wikidata: A Free Collaborative Knowledgebase’ (2014) 57(10) *Commun. ACM* 78.

would also be key; in this respect, much could be learned from proposed solutions to similar challenges in scientific data collaboration.¹⁰⁹⁰

Most data scientists are already used to working collaboratively online, through leading technologies in this space such as *Git*, *MediaWiki*, or *StackExchange*. Yet data scientists form only one part of the puzzle. As discussed both in this section and studied earlier in the thesis,¹⁰⁹¹ fairness issues can concern different parts of the modelling process, and as such viewpoints from others such as user interface developers, project managers, and decision subjects would likely be valid and useful. The technologies chosen should be clear and accessible to those who are not used to working in these virtual spaces, whilst incorporating the features and extensibility that more developed solutions bring. If they are not, they are likely to become exclusionary and not see the widespread adoption that would make them most useful.

It is not just modellers who can contribute information to this knowledge base. Quantitative and qualitative findings in the research literature that might be relevant to particular fields or data sources could be added. For example, considerable amounts of research exist on areas such as financial literacy, recidivism or child protection which are carried out with the aims of improving their fields, but not directly to make or inform decision support or decision-making. These forms of evidence could be used to directly inform model structure, or to inform in-data analysis and search for ethical issues and concerns. Many of these pieces of evidence are currently hard to locate—they are published across disciplines, behind paywalls, or with research questions that do not make clear the correlations that the research also unearths. In the medium term, text mining and natural-language processing might help populate such a database semi-automatically.

DADM/FATML methods, given their own technical opacity to laypersons, come with their own issues of transparency and legitimacy. As already discussed, individuals are, under the General Data Protection Regulation, entitled to know when automated processing of their personal data is occurring, and for what purposes, although there are important practical caveats regarding these rights.¹⁰⁹² Yet for them to understand the potential harms that could accrue to them by consenting is much trickier. Both they and trusted independent third parties usually lack the source data for investigative purposes. Even if they had it, it is unclear that it would be hugely useful or revealing given the rapidly changing nature of these datasets and the patterns within and the ample possibilities for data linkage that usually exist. Yet what they are (usually)

¹⁰⁹⁰ P Missier and others, ‘Linking multiple workflow provenance traces for interoperable collaborative science’ in *The 5th Workshop on Workflows in Support of Large-Scale Science* (IEEE 2010) DOI: 10/dsxdf6.

¹⁰⁹¹ Section 4.

¹⁰⁹² See section 2; see also Edwards and Veale, ‘Slave to the Algorithm?’ (n 79).

interested in is not the data themselves, but the potentially problematic patterns the data support. An evidence base might help individuals or organisations understand what insights are held in different forms of data.

5.2.2.2. Confounders

The proposal is largely grounded on the idea that organisations would be willing to spend time and money on cooperating to create a common resource. Primarily, this is a collective action problem, as there are great incentives to free ride and let others provide the information, which could result in non-provision.¹⁰⁹³ This is compounded by intellectual property concerns. If insights from data are viewed through an intellectual property (IP) or a trade secrets lens, this could make organisations reticent to share.

Yet sharing of data for ethical purposes between firms is far from unheard of, particularly in other sectors facing similarly tricky societal challenges. Social and environmental issues in the global clothing sector are pervasive due to uncertainties around the environmental impact of processes, materials and chemicals, and uncertainties in the on-the-ground production systems characterised by multi-layered subcontracting. The Sustainable Apparel Coalition (SAC) emerged as a data-sharing body in 2010, now with over 180 members representing well over a third of all clothing and footwear sold on the planet. Together with the US Environmental Protection Agency (EPA), and with several large data donations and collection projects involving members, they have been developing the open-source *Higg Index* to give designers tools to better and more rigorously anticipate potential products’ sustainability further upstream. In some ways, withholding data about ethical concerns and potentially salient social issues could itself be seen as a controversial, reputational risk.

Furthermore, the institutional field of the technology sector does not seem unamenable to this form of cooperation. Institutional fields create like-minded communities of practice through three main mechanisms – coercive pressure, where influence from actors or actants enforces homogeneity; mimetic pressures, which stem from standard, imitative responses to uncertainty; and normative pressures, which stem from how a field coalesces and becomes professionalised.¹⁰⁹⁴ Some promising normative pressures can be seen across the machine learning modelling field that give hope for this – communities of voluntary support on question–answer networks such as *Cross Validated* (which themselves support mimetic pressures); pro-bono data sci-

¹⁰⁹³ See generally, although contested, Olson (n 788).

¹⁰⁹⁴ Paul J DiMaggio and Walter W Powell, ‘The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields’ (1983) 48(2) *American Sociological Review* 147.

ence for non-profits on the weekends through growing organisations like *DataKind*; virtual discussions and events from field leaders on */r/MachineLearning* and *Quora*; expectations of contributions to open source software, to name a few. Proposed coercive pressures, such as professional bodies, charters or certification for data scientists might also play a role here in the future.

Identifying and creating databases of ‘good’ or ‘best’ practices is a common but also a problematic policy approach to complex socio-technical challenges. This approach can mislead, as practices are usually assumed to lead to good outcomes rather than being treated as hypotheses subject to serious monitoring and evaluation. Even where evidence suggests good practices work in one context, they may fail elsewhere.¹⁰⁹⁵ Instead of prescribing ‘good practice’, a database of experiences could serve a more exploratory function. Several organisations are well positioned to start or collaborate on such initiatives: private think-tanks such as *Data & Society* or *AI NOW* in the US, existing, new and proposed bodies such as potential ‘third parties’ outlined above in section 5.2.1.3, or one of many interdisciplinary collaboratives attempting to meld computer science and social science in universities across the world. It might also connect individuals facing similar challenges across the globe, creating creative, discussion-enabling support networks that help like-minded individuals share advice, strategies and even code to tackle the trickiest challenges together.

5.2.3. Exploratory fairness analysis

The situations above assume that information on protected characteristics are either possible to obtain, or available in parallel cases. Yet there may be situations where such data is restrictively difficult to obtain at all. Ambient computing, for example, judges people based on rather disembodied and abstracted features that environmental sensors can pick up, rather than through a data-entry method. Yet these systems might also exhibit fairness concerns; fairness concerns which might be particularly tricky to deal with.

These situations, where the protected data are not known, pose a difficult challenge for computational fairness tools. Yet I propose that there are concrete methods for these issues that while imperfect, could prove useful practices to both explore and develop in the future.

¹⁰⁹⁵ Nancy Cartwright and Jeremy Hardie, *Evidence-Based Policy: A practical guide to doing it better* (Oxford University Press 2012).

5.2.3.1. With unsupervised learning

Before building the model, data can be examined for patterns that might lead to bias. Exploratory data analysis is a core part of data analysis, but teaching, research and practice into it has been historically marginalised.¹⁰⁹⁶ Results of previous research, such as *DCUBE-GUI* or *D-Explorer*, have shown how visual tools might help with the understanding of potentially discriminatory patterns in datasets,¹⁰⁹⁷ even for novice users.¹⁰⁹⁸ Still, as with other methods, these tools broadly come with the assumption that the sensitive characteristics are available in the dataset, which I have argued is often unrealistic.

If we assume that immediately sensitive data are unavailable, simply understanding the correlations in the dataset is of less use. Instead, the exploratory challenge can be seen primarily an unsupervised learning problem. Unsupervised learning attempts to draw out and formalise hidden structure in datasets. Through unsupervised learning, we can hope to build an idea of the structure of correlations within data. As we do not have the sensitive characteristics, confirmatory analysis is difficult. This does not mean there is nothing to be done. Exploratory data analysis has much to contribute in the building of hypotheses and the directing of future data and evidence collection as part of a broader process of due diligence.

A relevant subset of unsupervised learning methods I zoom in on here attempt to understand dataset structure through estimating latent variables that appear to be present. Some methods, such as principal component analysis (PCA), try to create a lower dimensional version of the data that captures as much variance as possible with a smaller number of variables. Some social science methods such as Q-methodology use this approach to try and pick up latent dimensions such as subjective viewpoints.¹⁰⁹⁹ Other methods, such as Gaussian mixture models, assume that datasets are generated from several different Gaussian distributions, and attempt to locate and model these clusters.

These forms of analysis can be used to build hypotheses about fairness in datasets. For example, upon clustering or identifying subgroups within a dataset (which may or may not be related to any protected characteristics), these groups can be qual-

¹⁰⁹⁶ See John W Tukey, ‘We need both exploratory and confirmatory’ (1980) 34(1) *Am. Stat.* 23; John T Behrens, ‘Principles and procedures of exploratory data analysis’ (1997) 2(2) *Psychol. Methods* 131.

¹⁰⁹⁷ Bo Gao and Bettina Berendt, ‘Visual data mining for higher-level patterns: discrimination-aware data mining and beyond’ in *Proceedings of the 20th Machine Learning conference of Belgium and The Netherlands (Benelearn 2011)* (2011) (<https://perma.cc/E7R6-LYQK>); Bo Gao, ‘Exploratory Visualization Design Towards Online Social Network Privacy and Data Literacy’ (PhD, KU Leuven 2015).

¹⁰⁹⁸ Bettina Berendt and Sören Preibusch, ‘Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence’ (2014) 22(2) *Artif. Intell. & Law* 175 DOI: 10/gddxpd.

¹⁰⁹⁹ Bruce McKeown and Dan B Thomas, *Q Methodology* (SAGE 2013).

itatively examined, described and characterised. Experimental and sampling techniques might be used to gain more contextual information about the individuals in these clusters—for example, if their sensed or captured behaviour correlates with any sociodemographic attributes. These clusters can be used before or during the model building process to understand performance on different subgroups present in the data.

5.2.3.2. With interpretable models

A second approach to in-data analysis without access to protected characteristics examines trained models, rather than the input data alone. Once models have been trained, even complex models, there are several methods that are available for trying to understand their core logics in human-interpretable ways.

The literature on understanding models such as neural networks has traditionally distinguished between decompositional interpretation and pedagogical interpretation.¹¹⁰⁰ Decompositional approaches focus on how to represent patterns in data in a way that both optimises predictive performance whilst the internal logics remain semantically understandable to designers. Proponents of pedagogical systems on the other hand noted that it was difficult to get a semantically interpretable logic from models such as neural networks—although some do try.¹¹⁰¹ The tactic they have adopted, which is broadly the domain of most current research in interpreting complex systems, is to see the interpretation as a separate optimisation problem to be considered.

The concept of pedagogically interpretable models is relatively simple to explain. The basic idea is to wrap a complex model with a simpler one, which through querying the more complex model like an oracle, can estimate its core logics. Candidates include logistic regression or decision trees. Increasingly, proposals for the analysis of more complex models acknowledge that the gap between the logics that can be represented by the simpler model and the logics latent in a more complex model are too vast to translate appropriately. Image recognition is a case in point. Instead, proposals in this area have tried to estimate the logics that locally surround a given input vector—such as an image—to understand why it was classified as it was.¹¹⁰²

¹¹⁰⁰ Andrews, Diederich and Tickle (n 180); Tickle, Andrews, Golea and Diederich (n 180). Pedagogical interpretation has recently been described as ‘model-agnostic’ interpretation. See Ribeiro, Singh and Guestrin, ‘Why should I trust you?: Explaining the predictions of any classifier’ (n 207).

¹¹⁰¹ Y Jin, B Sendhoff and E Körner, ‘Simultaneous Generation of Accurate and Interpretable Neural Network Classifiers’ in Y Jin (ed), *Multi-objective machine learning* (2006) DOI: 10/frp7p6.

¹¹⁰² Ribeiro, Singh and Guestrin, ‘Why should I trust you?: Explaining the predictions of any classifier’ (n 207). This method is not exclusive to pedagogical methods, and some recent work has shown how decompositional methods, which use components of model structure rather than just treating it like a

5. Unpacking a tension: ‘Debiasing’, privately

Exploratory fairness analysts might manually examine mechanisms behind a model’s core logics and ask if they made sense. Specifically, analysts might wish to consider whether they would be happy publishing such information behind a model, or whether the public might take issue with the way and reasons behind decisions being made as they were. Some recent research that has highlighted gender bias in word embedding systems, which place words in relation to each other in high dimensional spaces to attempt to map different dimensions of their meaning, has gathered attention: and the methods of bias identification in this area are related to what has been discussed here.¹¹⁰³ Future work should tangibly explore whether meaningful and relevant information about datasets or models known to be somehow biased can be discerned through this type of analysis.

5.2.4. Discussion

The three distinct approaches I have outlined in this section point to three possible avenues for exploration in the research and practice of fairer machine learning. Each of them is suited for different purposes.

The third-party approach, where another organisation holds sensitive characteristics that they use to detect and potentially mitigate discrimination from data and models, is primarily useful where trust in the organisation interested in model building is low, or potential reputational risk is high. Insurance or hiring seem like prime cases here, particularly as they are areas historically associated with bias over protected variables. A challenge with this approach is that it is not easy to set up in low-resourced situations, or unilaterally.

The collaborative knowledge base approach, where linked databases featuring fairness issues noted and experienced by global researchers and practitioners, could be useful in a broad array of situations. It might provide benefit where general uncertainty is acute, risk assessment must be undertaken pre-emptively, or risks are complex, changing and sociotechnical. Yet this requires a change of mindset. Organisations involved in modelling should overcome a reluctance to openly discuss their models, and will need to dedicate time and money to give to as well as take from such a shared resource. Anonymous contributions could work as a model, but issues of who verifies provenance of the information given, and how easily it is to re-identify organisations based on modelling purpose would abound.

The exploratory approach requires the least organisational set-up, as it can be un-

black box, also display strong promise in this space. See Montavon, Lapuschkin, Binder, Samek and Müller (n 207).

¹¹⁰³ Caliskan, Bryson and Narayanan (n 119); Bolukbasi, Chang, Zou, Saligrama and Kalai (n 118).

dertaken unilaterally on data where sensitive characteristics are not held. Yet while this approach enables the construction of questions and the probing of certain types of anomalous or potentially problematic patterns in the data, on its own it provides by far the least assurance that fairness issues have been comprehensively identified, assessed and mitigated. Further work should seek to formalise methods of exploring data for these kinds of patterns, and test modellers and processes for their efficacy in identifying a range of synthetically induced issues.

There are, unsurprisingly, limits to the effectiveness of technological or managerial fixes to contested concepts such as fairness. Unsupervised learning is particularly challenging to evaluate fairness upon, given that groups discovered are latent, although there has been some recent work beginning to explore this space.¹¹⁰⁴ Understanding fairness by demographic will also be hard to grasp when those demographics are latent—such as treating individuals holding particular political views similarly in regards to moderating content online.¹¹⁰⁵ More importantly, even though the three approaches I outline deal with different levels of formality and different ways of understanding or conceiving fairness, they all remain broadly centred on the software artefacts themselves. I do not suggest here that either these approaches or the broad mindsets that underpin them are sufficient for understanding equity or mitigating discrimination in a digital age. I do however, tentatively suggest that where these software artefacts are used to make and support decisions, tackling technical aspects of these issues is likely a necessary piece of the puzzle—neither more nor less important than others, such as organisational culture, social methods of oversight, or decisions about the intention or direction of deployment. It is also important to draw attention to larger challenges with predictive systems: that they might not achieve social or policy goals at all by their nature,¹¹⁰⁶ or that fairness might not be the most relevant issue as much as ideas of stigmatisation, over-surveillance, or the devaluing of particular cultural notions, such as family units.¹¹⁰⁷ Where there are inherent conflicting interests between organisations deploying such systems and those affected by them, co-operation may not be feasible or desirable; affected groups may instead be drawn (understandably) to more adversarial forms of resistance and political action.¹¹⁰⁸

¹¹⁰⁴ Silvio Lattanzi Flavio Chierichetti Ravi Kumar and Sergei Vassilvitskii, ‘Fair Clustering Through Fairlets’ in *Presented as a talk at the 4th Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT/ML 2017), Halifax, Nova Scotia* (2017).

¹¹⁰⁵ Binns, Veale, Van Kleek and Shadbolt (n 1076).

¹¹⁰⁶ Harcourt (n 228).

¹¹⁰⁷ Anton Blank, Fiona Cram, Tim Dare, Barry Smith and Rhema Vaithianathan, *Ethical issues for Maori in predictive risk modelling to identify new-born children who are at high risk of future maltreatment* (, Government of New Zealand 2015) (<https://perma.cc/EWA7-VL4C>).

¹¹⁰⁸ See eg Finn Brunton and Helen Nissenbaum, *Obfuscation: A User’s Guide for Privacy and Protest* (The MIT Press 2015); John Danaher, ‘The threat of algocracy: Reality, resistance and accommodation’ (2016) 29 *Philosophy & Technology* 245 DOI: 10/gddv8k; David Lyon, ‘Resisting surveillance’ in Sean Hier and

Research directions These three proposals illustrate how alternative institutional set-ups and ways of knowing might help in the governance of fairness in the context of machine learning. It focusses on one identified practical constraint—the absence of sensitive data. Each approach introduces limitations, caveats and provides few guarantees of performance. This might irritate researchers in this space, yet it reflects the messy reality of many contemporary on-the-ground situations.

There are opportunities amidst the constraints. The practical limitations of fairness-improving approaches, including these three, will only become apparent upon their introduction and reflexive study within real-world settings. In particular, the second and third suggestions, concerning knowledge bases and exploratory analyses, are not amenable to the sort of mathematical guarantees that the discrimination-aware data mining literatures may find comforting. In these situation, process evaluation is much more important than outcome evaluation. Understanding the questions and challenges that these methods do (or do not) address during the real building, deployment and management of predictive systems is key here. Little work has been done in this space, and this should strongly increase. It is often unrealistic to assume mathematically sound ‘debiasing’ on-the-ground is possible, and this means it is often unhelpful to apply the validity conditions of traditional research in statistics and computer science to discrimination-aware machine learning. New technologies of this type should be at least partially assessed on the extent of new capabilities for responsible practices they afford practitioners—a difficult, transdisciplinary and heavily value-laden task, but a very necessary one.

Without this dimension, designed tools are likely to stumble in surprising and even mundane ways, which will affect their ability to deal with unfairness and discrimination in the wild. It seems unlikely that statistical guarantees of fairness will translate smoothly to individuals feeling that decisions about them were made fairly—something as much a result of process as of outcome.¹¹⁰⁹ Researchers working in this space should trial their proposed solutions, monitoring their implementation using rich and rigorous qualitative methods such as ethnography and action research, and feed findings from this back into tool revision and rethinking. To adequately address fairness in the context of machine learning, researchers and practitioners working towards ‘fairer’ machine learning need to recognise that this is not just an abstract constrained optimisation problem. It is a messy, contextually-embedded and necessarily socio-technical one, and needs to be treated as such. This requires technical scholars to better grasp the social challenges and contexts; but also for social scholars to grapple

Joshua Greenberg (eds), *The Surveillance Studies Reader* (McGraw-Hill 2007).

¹¹⁰⁹ On how explanation facilities fail to transfer smoothly into notions of procedural justice, see Binns, Van Kleek, Veale, Lyngs, Zhao and Shadbolt (n 212).

more rigorously with the technical proposals placed on the table, and to ensure that critiques with operational implications reach the ears of the computing community.

Part IV.

Joining the Dots

6. Synthesis and Recommendations

This thesis has treated a range of areas around the governance of machine learning. To recap, the main research questions it asked, using a variety of approaches, were:

1. To what extent might the current legal regime—European data protection law in particular—serve to successfully regulate harms from algorithmic systems? (*RQ1*, chapter 2, *The Law of Machine Learning?*)
2. Do practices and technologies surrounding machine learning stress this legal framework, and if so, where, why and how? (*RQ2*, chapter 3, *Data Protection's Lines, Blurred by Machine Learning*)
3. How congruent are the assumptions and framings of contemporary computational tools designed to tackle social issues in algorithmic systems with real-world environments and constraints? (*RQ3*, chapters 4, *Coping with Value(s) in Public Sector Machine Learning*; 5, *Unpacking a tension: 'Debiasing', privately*)
4. What practical actions might help society better govern machine learning systems? (*RQ4*, all chapters)

Firstly, I examined the motivating *algorithmic war-stories* that have captured the imaginations of researchers, the media, and at least a concerned portion of the public. Instead of defining a canonical harm or set of harms in a parsimonious framework, the concerns these war stories raised were taken as signs of the messy issues that *algorithmic systems*, particularly those integrating machine learning, might catalyse or give rise to. Following that, a narrative literature review of some of the technological fixes that have become popular topics of research was undertaken, placing Explaining, Debiasing and Accounting in the context of wider research and history.

I then pivoted from that to one of the main governance regimes thought to be a candidate for controlling machine learning that matters—European data protection law, the General Data Protection Regulation 2016 (GDPR) in particular. I outlined the provisions that appeared clear candidates for addressing these systems due to more specifically regulating algorithmic systems, such as the automated decision provisions

and information rights, and analysed them in the context of the technical characteristics of machine learning. The main barrier to the successful application of these rights to govern machine learning to the extent that some hope lies in their individualistic focus, which fosters a *transparency fallacy* to rival the failure of notice-and-content as providing meaningful choice to users, and the weak collective provisions which are needed to fill the vacuum between overburdened data subjects and outgunned regulators. Assumptions underpinning the theories of discovery, user agency and enforcement in the regulation might seem realistic on paper, but history indicates they are likely to be found wanting when meeting society.

However, I identify and analyse at least three areas in the GDPR that come with promise to mitigate this individualism–purpose limitation, data protection impact assessments, and certification systems. While these are also rife with unclarity, and often contain loopholes in need of plugging, they might provide some initial tools require to put the data subject’s rights and freedoms at the heart of machine learning without unduly responsabilising them in the process.

While promising tools exist, the thesis then considered whether, in light of machine learning systems, the framework *itself* was up to scratch. Data protection is premised on a range of underlying, implicit assumptions regarding the nature of the phenomena it attempts to govern using its ‘regulatory mixtape’ of tools, and in chapter 3, I argued that machine learning technologies and connected practices can challenge many of these premises. Firstly, there is an assumption that wherever individuals might be targeted with data-driven services or affected by machine learning, they will be easy enough to identify in datasets by the controller in order to exercise rights of control over this processing. Many practices around machine learning and the architecture of data systems challenge this. Secondly, there is an assumption that machine learning models themselves are software trained and queried with personal data, but itself distinct from it. This assumption appears to be faulty too, in light of recent evidence around attacks designed to recover personal data from trained systems. Lastly, there is a broad set of tensions around the distinction between *sensitive* (or special category) personal data, and ‘ordinary’ personal data. While this distinction is familiar in the algorithmic war stories—such as the *Target* war story, where loyalty card records were (allegedly) transformed into an estimate of the protected characteristic of pregnancy—this takes on quite a different flavour where quite radical data transformation occurs. The worked example used in this thesis concerns automated lipreading systems that transform video into an estimated transcript. The sensitivity of transcripts of speech is difficult to assess using the ‘categories’ of data that data protection uses to heighten protection, as both the content, implications and manner of speech might betray sens-

itive data within the meaning of the GDPR, such as health status or political opinion. Yet speech *does* reveal such categories, albeit not in the way that a census record does, and applying the same regime to speech as to databases of individuals consequently seems incongruent and creates practical governance challenges.

The thesis then moved from looking to tensions between technology and law to tensions between technology and practice. To do so, 27 public sector machine learning practitioners were interviewed around how they coped *today* with the value-laden aspects of the machine learning systems they were deploying or looking to deploy. This chapter highlighted tensions between the technical fixes outlined and reviewed in section 1.6 Computing to the Rescue? and the experience of practitioners in real, high-stakes settings—where machine learning truly matters. Again, when the assumptions of well-meaning researchers hit real-world scenarios, they appear incongruent in ways that are not easily predictable in advance.

Potential avenues to deal with this situation were analysed, taking one of these tensions between research and practice as an example: the lack of consideration that sensitive data is *required* for debiasing, yet using such data might cause privacy or trust issues in turn. These included finding alternative arrangements, particularly cryptographically, to access or perform computations for detecting and mitigating unwanted patterns; catalysing a ‘knowledge base’ to draw on times when such data *was* present in parallel domains; or engaging in ‘exploratory fairness analysis’ to find patterns and structure in data to justify targeted data collection and hypothesis testing. These approaches point to a much wider array of frames than currently seen through the fairness literature, which largely remains preoccupied with constrained optimisation problems. However, the epistemological shift would, in general, necessitate a move away from seeking formal mathematical proof that a given issue was solved to a procedural focus on seeing that a range of issues were examined and steps taken to mitigate them to the best abilities of all involved.

I now present some themes distilled from the chapters that preceded. While chapters (and subsections within chapter 3) can and do stand alone as pieces of work, their synthesis and emergent cross-cutting lessons form important overarching takeaways. Because this work is designed to inform and support policy-making, I have—perhaps unusually for a PhD thesis—explicitly drawn out recommendations for researchers, regulators/legislators and users (of machine learning tools) within these themes. This sits in the spirit of the use-inspired and young department this thesis has been written in: seeking to create academic work oriented and driven by accessibility

and practical relevance. The themes presented here are not exhaustive representations of the work above, but represent areas of overlap across it.

6.1. Going beyond canonical ‘decisions’

This thesis illustrated how the governance of machine learning systems and the salient legal provisions in data protection law so far have primarily focussed upon certain canonical decisions (sections 2.2.1.2), such as recruitment or recidivism. These typically have clear analogues in previous choices made by individuals or within organisations, and thus are instantly recognisable as a decision that has or could be ‘de-humanised’. Concerns around this flavour of decision have been especially motivated by the algorithmic war-stories that have served to organise the field (section 1.4), and many issues, such as the retrenchment or replication of discrimination on the basis of groups such as sex or race are indeed severe and should not be downplayed.

What this thesis advocates instead is the ‘up-playing’ of several other processes or measures machine learning contributes to, which may or may not be easily classified as a ‘decision’ in the canonical sense of having a clear, human-determined analogue. Section 2.2.1.2 highlighted that upstream choices, such as applying or generating a risk score that is then widely propagated, might be the most appropriate location to identify a consequential decision, particularly if the downstream impacts all have minor but additive effects on individuals or groups they constitute. Section 3.1.5.3 argued that other upstream choices related to machine learning might constitute significant decisions worthy of heightened governance, such as the choice to remove identifiers from an inherently identifiable training dataset, which effectively limits data subjects’ ability to draw upon rights including erasure, access and objection. These points of action become more interconnected in a world where data flows and machine learning pipelines become increasingly prominent, and accountability becomes blurred as a result. Where to draw the system boundaries remains an important and value-laden task: the wider they are, the less focus is placed on the software alone, and the more broader and highly important questions of power and political economy enter the fray. Suggesting that the ‘last-hurdle’ of decision-making is the point to hold to account the most risks *avoiding* tackling some of the systemic issues at play when machine learning matters, and those involved in shaping governance of this space should recognise that this has high potential as a lobbying tactic to help actors dodge liability or accountability.

Recommendation 1 (Regulators) *The notion of a ‘decision’ related to machine learning should be explicitly recognised to include upstream choices with potentially distributed or*

delayed downstream impacts. Controllers should be required to analyse the systems they operate in to identify these influential points and undertake appropriate obligations.

6.2. Building and scrutinising process

The work in this thesis has emphasised the broader role of *process* as a counterpoint to the usual focus on specific decision points described above.

In both chapters 2 and 3, the potential use of data protection impact assessment (DPIA) in making trade-offs and choices was highlighted. Yet, except in public sector situations (where they could be FOI'd) or where a private company has volunteered them, these will generally not be accessible to the public, limiting their use in scrutinising choices or holding powerful actors to account. There is a risk that such secret documents may turn into a box-ticking process as a result, serving mainly to minimise regulatory penalty should an investigation be carried out. Yet, as a prototype of a policy instrument that could be later strengthened or developed, they hold promise for both expansion to other considerations (as the rights and freedoms that must be assessed are general and not linked to data protection alone) and for creating future scrutiny systems around.

Recommendation 2 (Regulators and Users) *As can be read into the GDPR, DPIAs should consider all relevant rights and freedoms associated to personal data processing. These documents should be released to the public.*

As data sources are mixed up, and as systems change through processes of feedback and concept drift, there is not just a statistical challenge, but an organisational one. As shown particularly in sections 4.6.1.4 and 4.6.1.3, different actors in the data pipeline are likely to have different motivators, and these divergent interests can quickly create surprising failure and friction. Transparency in the form of explanations have previously focussed on decision-support users and now increasingly are focussed on those decisions affect (see section 1.6.1), but I believe that in complex sociotechnical systems with significant algorithmic components, we need to consider how explanations function *within* organisations rather than just outside of them. Without a concerted focus on how this can be done, it seems unlikely the the preponderance of ethics boards, chief privacy or information officers popping up in organisations worldwide are likely to provide much input beyond legitimising of systems inside their organisations on paper.

Recommendation 3 (Research) *Researchers should go beyond explanations aimed at decision-makers or decision-subjects to consider how to best provide useful transparency to*

6. Synthesis and Recommendations

those with other functions within organisations like firms or public bodies, such as those collecting or determining data collection processes.

A focus on process also spurs further consideration of over-reliance, under-reliance and augmentation of the results of algorithmic systems. This thesis explored the notion of what is considered ‘solely’ automated under the GDPR, which pointed to tricky organisational aspects needed to give decision-support users the confidence to disagree with the machine (section 2.2.1.1). The empirical work in this thesis pointed to the difficulties in both establishing these processes of responsible decision-support but also, importantly, in documenting, maintaining and transferring them across contexts and over time (section 4.6.2.1). Such practices of documentation—a ‘what works’ of sorts—also appeared useful in other contexts, such as in knowledge-bases to understand what kind of biases could appear in different types of datasets in situations of limited information (section 5.2.2). Yet to do so will likely require context-specific capacity, as over- and under-reliance appear to be situation and application specific (section 4.6.1.2).

Recommendation 4 (Users) *Organisations seeking to mediate machine learning decisions with human input must pay careful attention to ensuring such input is meaningful, and importantly, to attempt to ensure that the social, technical and organisational structures are durable across time and contexts.*

Recommendation 5 (Regulation) *Rule-setting organisations should implement requirements for documenting (potentially in DPIAs), monitoring and evaluating human mediation and oversight and its supporting context in decision-support systems.*

Recommendation 6 (Research) *Researchers should seek to understand the conditions for rigorous human oversight of systems and how the quality of oversight can be monitored in an efficient and scalable manner.*

6.3. Anticipation needs

Another theme that can be drawn out of this thesis is the importance of an organisational capacity for *anticipation*. By this, I mean that it is difficult to place safeguards or to consider impact in both the way that the law and responsible use of the technology more generally demand without significant capacity to think about the future.

This was shown in several areas of the thesis. Firstly (section 2.2.1.1) it is hard for data controllers to understand what was a *significant* decision for the purposes of data protection law, potentially to an unexpected vulnerable subgroup or in a surprising

manner. This extends to understanding the issues that could mediate human under- or over-reliance on decision-support, and thus affect whether a provision fell inside or outside of Article 22, GDPR's provisions. These issues appeared again in the empirical work, since as just mentioned, how individuals did or did not trust systems appeared to be context dependent in ways we do not currently appear to understand well (section 4.6.1.2). It could be argued that without anticipation capacity, data controllers should be expected to find a lawful basis for many more automated decisions than might actually fall under Article 22, in order to provide what the CJEU describes as 'effective and complete protection'.¹¹¹⁰

Anticipation capacity is also required to limit the impact of unexpected failure modes in algorithmic systems. Empirical work highlighted unexpected errors: systematically erratic data subjects biasing models and gaming by decision support users and decision subjects, among other issues. (sections 4.6.1.4, 4.6.2.4). These failure modes result from the interaction of social and technical systems in unexpected ways. While many times they cannot be predicted, it might be that anticipating the types of failures that could occur, and the ways in which an algorithmic system could foster surprising results, could enable strategies to better assess or mitigate their impact.

Looking ahead is also important when considering how processing techniques and architectures interact with the provisions of data subjects' rights and freedoms. Section 3.1 described how removing identifiers from datasets, common when using these data to build machine learning models, can effectively shift risk onto data subjects whilst simultaneously denying them the tools to manage it, such as rights of access, erasure and objection. Such processing techniques, intentional or unintentional, emphasise the interaction between upstream choices and downstream autonomy and control: consequences that require effective anticipation capacity to cope with or envisage. Data protection by design (DPbD), while admirable and ostensibly very forward looking, comes with future trade-offs which can only be responsibly made by estimating their shape, form and impacts on different individuals and groups.

This thesis also draws attention to the shifting notion of sensitivity of data, how that is and can be transformed over time and in different contexts, and the need to preempt such transformations. The case study on automated lipreading systems (ALRSs) in section 3.3 drew attention to how infrastructure that has been established under certain conditions (eg that it could not capture audio) might find itself repurposed by machine learning without the necessary safeguards or congruent legal framework to land within. What is sensitive and what is not is challenging, but necessary, to anticipate in advance.

¹¹¹⁰ *Google Spain* (n 405) ¶ 34; *Wirtschaftsakademie* (n 403) ¶ 28.

Recommendation 7 (Research) *Researchers should adapt, develop and assess anticipation methods for social issues concerning machine learning.*

Recommendation 8 (Regulation) *Regulators should promote rigorous methodologies for anticipation, and for high-risk uses of machine learning, look for evidence of their use in DPIAs.*

There are significant challenges in pushing for better anticipation in this field, however. Perhaps the main one relates to scale and required capacity. Machine learning, like many digital systems, is characterised by its increasing asymmetry. Previously, profiling at a global scale was only within the reach of governments or multinationals. Now, *in extremis*, it could be deployed on cloud infrastructure from a £10 *Raspberry Pi* from someone's bedroom. Critical evaluation, monitoring, compliance, anticipation or foresight capacities have not scaled similarly. As smaller and smaller firms have more influence and global reach, regulating them and placing anticipatory capacity expectations upon them is likely to catalyse calls of overburdening or onerous regulation. These calls are likely correct—it is unrealistic to ask such firms to take on responsibilities they do not have the manpower, expertise or financing to undertake.

All in all, it begs the question that has proved difficult to ask—should firms this small be allowed to adopt such influential roles at all? And if not, how would they be restricted with minimal effect on individuals' online liberties?

6.4. Rethinking the role of groups

Another cross-cutting area this thesis has highlighted is the importance and varying roles of groups.

Groups are important, and I have argued, necessary additions to the legal framework that currently receive little consideration. Firstly, groups can be significantly affected by algorithmic systems in ways that individuals alone might not be. Data protection law, as it stands, emphasises individual impact, and as a result, fails to adequately address this issue (section 2.3.2). This is additionally emphasised when considering the gaps in governing models and their movements in the world beyond when they are being trained and queried in reference to a particular individual or decision. Looking through the lens of the groups illustrates reasons why users might collectively want to explore models, and section 3.2 demonstrated how they might legally be able to—albeit through an interpretation of the law that, undesirably, disconnects the triggering condition from impact or effect.

Recommendation 9 (Regulation) *Regulators should ensure that machine learning systems that might not affect an easily identifiable individual significantly but might, in aggregate, disadvantage a group, are sufficiently considered in relevant governance frameworks.*

Groups are also important, as they might provide pluralistic oversight. Particularly (although not only) where machine learning systems affect a group or class of individuals, not all of them may be in the same empowered position to utilise their rights, raise complaints or take data controllers to court. It could be said they very rarely would be. In these situations, it is critical that the vulnerability of these individuals does not act as a reason to deny them adequate protection under the law. As this thesis has examined, the collective provisions of the GDPR are not always adequate or effective in achieving this aim (section 2.3.2). However, as it has also examined, collective action by data subjects could be powerful in the context of machine learning systems, potentially creating situations of mass objection or erasure which could have impacts on business models and the use of deployed systems (section 3.2.5). Furthermore, the regulation completely lacks abilities for civil society organisations to engage in investigative action against powerful data controllers. While there are privacy concerns with extending this carelessly, as civil society organisations should clearly not be given access to individuals' personal data without consent in the way that regulators have the power to examine, middle grounds, such as receipt of aggregate information or examination of models in secure environments such as those used to facilitate the reuse of microdata by statistical agencies, could be envisaged.

Recommendation 10 (Regulation) *The UK government should make provisions for the collective provisions under Article 80(2), GDPR within the Data Protection Act .*

Recommendation 11 (Regulation) *Legislators should ensure that revisions of data protection law allow rights beyond those in Article 80(1) to be delegated, such as access, erasure and objection, and may wish to consider means by which civil society organisations have some access to consequential machine learning systems in private organisations for investigative purposes.*

Recommendation 12 (Research) *Researchers should consider how to collectively mobilise the right to access to better enable oversight of machine learning development processes whilst maintaining individual privacy. Privacy-preserving computation might be a good contender to help individuals better manage their data and combine it to enable aggregated oversight.*

Finally, the nature of groups should be reconsidered in light of the new ways that data-driven systems generate nuanced structure in the world. Data protection's approach to sensitive groups is largely visible through the 'special categories' of data in

Article 9. These include some traditional protected classes, such as ethnicity, religion or disability, as well as some non-traditional groups, such as individuals within a trade union or holding certain political opinions. However, this only scratches the surface of the constructed, latent groups which machine learning might target that differ strongly from protected characteristics—such as socioeconomic status—as well as a heightened focus on the combination of characteristics an intersectional lens would offer. Such groups might be difficult to discover—as analysed in chapter 5—or might be so heavily context specific to not be considered in cross-cutting legislation such as equality law or anti-discrimination. Google, for example, provides an API to predict the probability a particular downloader of a gaming application on their Android OS' *Play Store* will spend an amount within the 95th percentile or higher of a game's user base:¹¹¹¹ something forming a latent group of individuals with high vulnerability or susceptibility that are particularly commercially exploitable. In the empirical research section of the thesis, it was seen that actors did not always look to protected characteristics to understand groups they should avoid targeting unequally. For tax agencies for example, even distribution of investigations across the country was intrinsically important, as opposed to just being a concern due to acting as a proxy for some other characteristic such as ethnicity (section 4.6.2.3). Yet detecting and discussing these groups is not straightforward, and as almost the entirety of work on fairness in machine learning has focussed on canonical protected groups (see section 1.6.2), few technical practices exist for understanding groups not explicitly identified in the data (as section 5.2.3 suggests is required).

Recommendation 13 (Research) *Researchers should go beyond explicitly defined vulnerable groups to also consider latent groups, and develop techniques designed specifically to highlight them and determine their importance.*

Recommendation 14 (Regulation) *Legislators should consider what obligations be placed on users of machine learning systems to identify groups that may be disproportionately affected by a deployed system additional to those in anti-discrimination law.*

* * *

Machine learning matters, even if only because it scales existing sociotechnical challenges in impact, reach and tempo. It exacerbates, as many technologies do, issues of asymmetry of effect. Both small businesses and individual units of government can, in theory, achieve much greater reach and impact than previously, without

¹¹¹¹ Ronan Fahy, Joris van Hoboken and Nico van Eijk, 'Data Privacy, Transparency and the Data-Driven Transformation of Games to Services' in *2018 IEEE Games, Entertainment, Media Conference (GEM)* (IEEE 2018) DOI: 10/gfsvz6 141.

establishing an unwieldy bureaucracy. Upon first glance, this would appear to aid consistency, which we might consider an important aspect of procedural justice. On the other hand however, this consistency is consistency of a tool rather than its impact. Machine learning does not easily consider the differing vulnerability of the individuals and groups it is applied to.

Machine learning comes with a tempting offer for those interested in social justice. Scrutiny of imperfect, biased and often opaque *human-made decisions* has always been problematic. Yet machine learning systems seem at first glance to be crystallised single software objects which might be amenable to scrutiny if such oversight procedures can be established. The problem is that achieving such oversight is far more difficult in practice than it is on paper. Data and systems are proprietary and private. Processes and software have an ad hoc flavour, are generated or tailored in-house, and their value-laden components are glued together by crucial tacit knowledge. Few bodies are capable of providing meaningful oversight of complex, sociotechnical issues, and those that are capable are usually deluged. Assessment of social challenges can require complex methodologies still under development, or unavailable data.

The hope of legislators has hinged largely on limited individual transparency tools providing cheap and scalable oversight. This is a myth, and in my view, a dangerous one to rely on. Expecting vulnerable individuals affected most by single or cumulative automated decisions or measures to lead the charge in inspecting, analysing and assessing them is absurd in an environment where data protection regulators themselves appear overwhelmed and outgunned by the technologies, practices and scale of the challenge at hand.

The hope of computer scientists of automating regulation or oversight is also premature. Pre-built solutions for issues such as fairness or accountability cannot be pre-tailored for problem context and social processes that surround consequential systems. Even if tailorable, air-dropped tools are still dropped into an already framed problem. If issues exacerbated by machine learning emerge from the problem framing instead of or in addition to specific implementation, then technical fixes might fix little of consequence, and indeed doubling down on that framing might lock systems into undesirable states. This does not mean there is no room for those tools—but that they are no panacea. For them to be of specific use, they do however need to be tailored more to use case and context than they currently are, as toy problems and benchmark datasets differ strongly from messy institutional conditions on-the-ground.

Interdisciplinary theses usually end with stock calls for further or more intense interdisciplinarity. I want to end this thesis by calling for a *different* type of interdisciplinarity than the one that is emerging now around machine learning and society. This

field, broadly defined, merges different disciplines by default, but the way it does so could, and I feel should, be otherwise. Law, particularly older law, is distilled and de-contextualised, and cherrypicked by computer scientists, who look for parts of it that are able to be implemented within the accepted mode of knowledge-building within their subfield. This in turn is promoted as a solution, or at least a useful approach, for the societal challenges that machine learning exacerbates, and which are only beginning to be well-understood.

The order of this approach seems inherently muddled and divided. Instead of actors in this space wondering how they can take an action to mitigate the problem as they perceive it, I believe that the different disciplines and stakeholders should heighten their focus on how their domains can increase the capacity and reduce the barriers for society to collectively frame, organise and tackle these important issues. This capacity-building might not resemble some of the ‘silver bullets’ placed on the table today—a law of explanations, a system to scrub bias, a journalistic outfit to expose, name and shame. It might, instead, grease the wheels, provide support to and increase the analytic abilities of activists, regulators, data scientists to define their own strategies *in situ*. All disciplines can surely help with this challenge. By recognising they are each just part of a broader sociotechnical system characterised by tricky problems and stubborn power, work that seems like a modest stepping stone might have a much more transformative effect than might first have been imagined.

Part V.

In the Back

7. Appendices

7.1. Apple Correspondence

06-07-17

Dear Apple,

I would like to make an access request under the Data Protection Act for the recordings associated with my Siri identifier. As noted by the ICO and the Article 29 working party in relation to Wifi analytics in recent guidance, an identifier such as a hash or a random number is also considered personal data if it is not salted daily. Consequently, voice data that is reidentifiable to me (by virtue of being being able to delete it by turning off Siri, as Apple can link it to my device), even if not linkable directly to my Apple ID or name, is data which I have access rights over.

I would like all recordings associated to the Siri identifier linked to my personal device. Please let me know what information I need to provide to make this access request.

Michael Veale

24-07-18

Dear Mr Veale,

I am replying on behalf of Apple Distribution International, Ireland, which is the data controller for Apple customer personal data in the EU/EEA.

Thank you for your request in relation to Siri data. I would like to direct you to the information in relation to Siri which can be found on iOS Devices in the Settings app: please look under Siri - About Siri and Privacy. For ease of reference I have copied the text as an attachment to this mail. We also have privacy relevant information available on Siri here:

<https://www.apple.com/de/privacy/approach-to-privacy>

As you can see we have designed Siri to be privacy by default such that we cannot link Siri data, including audio files, with an identifiable user. We have no means to access or receive your Siri identifier from your device.

Kind regards,

7. Appendices

REDACTED

Apple Privacy

24-07-18

Dear REDACTED,

That information notes that the Siri explicitly stores a personal identifier – the name of an individual.

“Your device will also send Apple other information, such as your name and nickname”

Please provide details on how this name data is not able to be connected to the audio files or transcripts. If the name data can be connected to the audio data, it is clearly considered personal data under the Data Protection Directive, regardless of whether the Siri identifier is unable to be retrieved from a device by the user.

Michael Veale

03-08-18

Dear Mr Veale,

Thank you for this further response, we welcome your questions on this issue.

I think it is important to point out that we are not somehow seeking to evade complying with your access request. The opposite is the case. We take our meeting of user privacy rights very seriously. We work hard to produce products and services which put privacy first and Siri features prominently in that list.

Some of the assumptions made in your response do not reflect the technical engineering of Siri. As explained in more detail in our iOS Security White Paper (https://www.apple.com/business/docs/iOS_Security_Guide.pdf), your Siri name and nickname are chosen by you, the user. They are not tied to an Apple ID and are not verified. Siri data is keyed (indexed) by the random identifiers stored by your devices. It is therefore stored in a manner not practically searchable, and we have not built any tool that allows us to retrieve this data.

Because of the design of Siri, we are therefore not able to provide any responsive data for Siri in response to a subject access request or any other third party request, including those from law enforcement.

Kind regards,

REDACTED

Apple Privacy

03-08-18

Dear REDACTED,

Thank you for your response.

I appreciate that Apple is attempting a Privacy by Design approach to Siri data, and this is welcomed.

My main point of contention is that I am not convinced, from the information that Apple has published in the previous document and in the one you have just linked to, which I have read, that Siri data is not personal data under data protection law.

My core question (1) surrounds the connection of personal data able to 'single you out' to Siri voice data and transcripts.

The security White Paper notes the list of data that is sent to the server in a non transient way (i.e. not like the name, which is sent and erased in a short time period).

To facilitate Siri features, some of the user's information from the device is sent to the server. This includes information about the music library (song titles, artists, and playlists), the names of Reminders lists, and names and relationships that are defined in Contacts.

Much, if not all, of the data described here would almost certainly be considered personal data for the purpose of data protection. It is not hard to see how reidentification on the basis of names and relationships of contacts, for example, would be extraordinarily easy with public data.

My question is whether this data shares *any* common identifying variable with the Siri voice data.

If this does, then data protection rights, including the rights of access and portability, will apply to this data. As you will be aware, simply having not developed a tool to extract data does not exclude Apple from following these rights. The question surrounds the feasibility of designing such a tool. If this tool cannot be designed for mathematical reasons, please evidence .

In addition (2), voice transcripts themselves contain personally identifiable information divulged by the speaker. Please provide information about redaction techniques used to mitigate these data, if any are used.

Note that the point you made about a Siri name not being verified does not mean that it is not personal data. An individual providing valid personal data expects to rely on data protection rights and obligations.

Please note that I am not attempting to ask the impossible. If this data are truly not identifiable, then that is a satisfactory response. However, none of the documentation

7. Appendices

provided so far have convinced me that this is the case .

Best,

Michael Veale

16-08-17

Dear Mr Veale,

Thank you for your further email and I apologise for the slight delay in replying due to leave. As we understand it, your main unresolved question appears to relate to the questions of identifiability of the data once it is in our possession. While we note that you indicate that you remain unconvinced by our responses, we actually consider that the information that we have provided in fact fully answers the question of whether data collected from the use of Siri is considered personal data under the provisions of applicable data protection law, which in this case is the Irish Data Protection Acts 1988 & 2003.

Data which does not relate to or cannot be linked to an identifiable person is by definition not personal data. We understand, of course, that the question of whether a set of data can or cannot be linked to an identifiable individual must take account of all the potential data points collected which might render data that is otherwise anonymous to become identifiable and therefore within the law. It must equally also take account of the ability to retrieve information against those data points. The Irish Data Protection Acts defines this as “accessible according to specific criteria.”

Section 1(1) definitions: “relevant filing system’ means any set of information relating to individuals to the extent that, although the information is not processed by means of equipment operating automatically in response to instructions given for that purpose, the set is structured, either by reference to individuals or by reference to criteria relating to individuals, in such a way that specific information relating to a particular individual is readily accessible’

As we look ahead to the GDPR, it states:

Article 2 - Material scope

1. This Regulation applies to the processing of personal data wholly or partly by automated means and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system.

Article 4 - Definitions

(6) ‘filing system’ means any structured set of personal data which are accessible according to specific criteria, whether centralised, decentralised or dispersed on a functional or geographical basis;

It is therefore clear to us that the intent of the legislator was and will remain that in order for data to be considered to be personal data they must be stored in a filing system which is accessible according to specific criteria relating to the individual.

To summarise again, information collected from the use of Siri is not tied to an Apple ID and is not verified on our systems. Furthermore, the name data cannot be connected to the audio data because, although it is sent to Siri, it is stored in a manner that does not render it searchable according to specific criteria. Siri data is keyed (indexed) only by the random identifiers stored by users' devices. In short, we do not have a technical means to access or link the random Siri identifier to the user.

I hope that the above is of assistance.

Regards,

REDACTED

Apple Privacy

15-09-17

Dear REDACTED,

I have consulted with leading data protection academics and they agree that your latest response is absurd.

You appear to be arguing that there is no automated processing of Siri data by Apple.

I point you to the article you quote:

1. This Regulation applies to the processing of personal data wholly or partly by automated means and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system.

The test of whether or not data forms part of a filing system only applies to data that are processed other than by automated means. Siri data is processed by automated means. Therefore only first part of that sentence (before the 'and', italicised) applies. Your argument is nonsensical. Simply because a firm has not built a system to retrieve the data when it is clearly technically possible does not exempt it from DP rights. If this was true, no data controller would have to accommodate DP access requests. Please provide alternative justification.

Secondly, there is no requirement for data to be 'verified' in the DPA or the GDPR in order for rights to reply. I understand that there may be a need for me to verify myself before making the request, but DP rights apply regardless of verification tests undertaken by the data controller, Apple. If you know of such a requirement, please quote it from the legislation.

In addition, please address my previous questions: 1) whether contact data shares a common variable with Siri recordings and/or transcripts in the data structure; 2)

7. Appendices

whether redaction techniques for identifying aspects of Siri transcripts are applied to de-identify them, and if so what they are.

Best,
Michael Veale

10-10-17

Dear Mr Veale,

Thank you for your email of 15 September and your continued communications on this matter.

In every product and service that Apple engineers we take great care to protect our customers' privacy and keep their data secure. Our recently updated privacy website makes this commitment clear.

As we have told you, we are not able to satisfy your request because your Siri identifier does not identify you by name or relate to or reveal any other personally identifying information. The irony of your request being that to satisfy it, we would have to create a data collection tool with stable identifiers that we would regard as privacy unfriendly. We're pleased to have had the opportunity to explain all of our privacy by default features to you. We trust you will agree that Siri represents privacy by default where access is not technically possible and that you consider this complaint closed.

Regards,
REDACTED
Apple Privacy

18-06-18

Hello,

I am writing to request a copy of all Siri audio and transcript data associated with my Apple Macbook Pro under Article 15 of the General Data Protection Regulation 2016. I have attached a copy of a file from my device which I believe to contain the persistent Siri Identifiers generated by my device, which are used to index voice recordings and transcripts on your server for the purpose of increased personalisation of modelling. Please let me know if this is not the correct file, and I will retrieve the correct one from my computer. I would like these sent to me in electronic form.

Please note that there is no exemption for disproportionate effort in Article 15 of the GDPR as there was in the previous Irish Data Protection Act.

Best wishes,
Michael Veale

17-07-2018

Dear Mr. Veale,

I refer further to your attached request.

We note your continued intention to seek access to Siri audio and transcript data. We know also that you fully appreciate that Apple went to extreme engineering efforts to develop and maintain a truly industry leading voice assistant that contained the very best privacy protections for our users. We also understand that you are approaching this matter from a strong privacy preservation perspective.

In this spirit, we can confirm that the file in question that you provided does not include any relevant Siri identifiers. Even if it did contain such identifiers we have no means to authenticate that a Siri ID provided in this manner is in fact the ID associated with your device. Only authenticated devices do so, and for us to do otherwise would weaken the privacy protections of our design.

If you are not satisfied with our answer, you are within your rights to contact the ICO or Irish Data Protection Commissioner directly, and we would be happy to provide you with contact details for them if needed.

Finally, we do thank you for continuing to raise these matters. Thoughtful contacts such as those which we receive from you do provide a valuable input as we challenge ourselves everyday to provide the very best privacy protections for our users.

Kind regards,

REDACTED

Apple Privacy

7.2. Workshops and Conferences Attended

Name	Location	Date	Role
IRGC Risk Conference	London	Jan 2016	
NESTA Machine Learning in Government Workshop	London	Feb 2016	
Alan Turing Institute/Bartlett Future Cities Conference	London	Feb 2016	
Alan Turing Institute Algorithm Society Workshop	London	Feb 2016	
International Data Responsibility Conference	The Hague	Feb 2016	
Royal Academy of Engineering Securing Critical Infrastructure Workshop	London	Mar 2016	

continued.....

7. Appendices

.....continuation

Name	Location	Date	Role
Alan Turing Institute Responsible Research and Innovation Workshop	London	Mar 2016	
Knowing Algorithms Workshop, Lancaster University	London	Mar 2016	Speaker
Responsible Research and Innovation Tools Workshop	London	Apr 2016	
Cybersecurity Workshop, FCO	London	May 2016	
Superhuman: Human and AI in Synergy, Microsoft Research Cambridge	Cambridge	May 2016	Speaker
Big Data: Does Size Matter Launch	London	May 2016	
UCL Crime Policing Citizenship Conference	London	Jun 2016	
HYPERCAT Summit	London	Jun 2016	
Science and Democracy Network Annual Meeting	London	Jun 2016	
Technology Management and Policy Conference, University of Cambridge	London	Jun 2016	Speaker
Algorithms and Society Summer School	Stockholm, SE	Jul 2016	
Interdisciplinary Summer School on Privacy	Nijmegen, NL	Jul 2016	
Projects by IF Seminar	London	Jul 2016	Speaker
Alan Turing Institute Workshop on Machine Learning	Edinburgh	Aug 2016	
MCCRC Machine Learning: Technology Law and Policy Symposium	Cambridge	Sep 2016	
EIT Pit Stop: Digital Local Government (Digital Catapult)	London	Sep 2016	
Data for Policy 2016	Cambridge	Sep 2016	Poster
University of Southampton Text and Data Mining and the Law Workshop	So'ton	Sep 2016	
DataKind DataDive	London	Sep 2016	
ANYWARE: Location Data and the Law, VUB	Brussels, BE	Oct 2016	
AI: Risks and Opportunities, IET	London	Oct 2016	
Launch: Centre for Future Crime	London	Oct 2016	
Mozilla Festival 2016	London	Oct 2016	
Data Transparency Lab Workshop	New York	Nov 2016	
FAT/ML 2016	New York	Nov 2016	
DAT 2016	New York	Nov 2016	
Ethics of Machine Learning in Professional Practice, University of Cambridge	Cambridge	Nov 2016	Speaker
Human Centred Computing Workshop, University of Oxford	Oxford	Nov 2016	Speaker
Knowledge and Innovation Network, University of Warwick	Warwick	Dec 2016	Speaker

continued.....

7.2. Workshops and Conferences Attended

.....continuation

Name	Location	Date	Role
The Human Use of Machine Learning Workshop, Ca'Foscari University	Venice	Dec 2016	Speaker
UnBias Stakeholder Workshop	London	Feb 2017	
Algorithm Workshop, University of Strathclyde	Glasgow	Feb 2017	Speaker
Internet of the Future Workshop, University of Cambridge	Cambridge	Feb 2017	
AI For Social Good, Waseda University	Tokyo	Mar 2017	Speaker
Ethics, Computer Systems and the Professions	London	Mar 2017	
Jill Dando Institute Secure Data Centre Launch	London	Mar 2017	
Responsible Research and Innovation in ICT, St Cross College	Oxford	Mar 2017	Speaker
Politics of Algorithms, ULB	Oxford	Mar 2017	Speaker
Royal Statistical Society Algorithms in Decision-Making Roundtable	London	Apr 2017	Speaker
Modern Slavery and Policing Conference, Cumberland Lodge	Egham	Apr 2017	Facilitator
Realising the Potential of AI in Financial Services, techUK	London	Apr 2017	Speaker
Public Sector Machine Learning Workshop, King's College London	London	Apr 2017	Speaker
How can data scientists build unbiased systems when society is full of bias, DataKind Seminar	London	Apr 2017	Speaker
4th Winchester Conference on Trust, Risk, Information and the Law (TRILCon 2017), University of Winchester	Winchester	May 2017	Speaker
Big Data: New Challenges for Law and Ethics, University of Ljubljana	Ljubljana, SI	May 2017	Speaker
Privacy Law Scholars Conference (PLSC), UC Berkeley	Berkeley, US	Jun 2017	Speaker
Cambridge Technology Ventures Conference	Cambridge	Jun 2017	Speaker
CognitionX Conference	London	Jun 2017	Speaker
LSE CARR/KCL Algorithmic Regulation Workshop	London	Jul 2017	
AI and the Law, BBC	London	Jul 2017	Speaker
Machine Learning and Policy, TNO	The Hague	Jul 2017	Speaker
ACM KDD 2017	Halifax, CA	Aug 2017	
FAT/ML 2017	Halifax, CA	Aug 2017	Speaker
Amnesty International Machine Learning Workshop	London	Sep 2017	Speaker
Data for Policy 2017	London	Sep 2017	
MCCRC Blockchain: Hype meets the Long Arm of the Law, Cumberland Lodge	Egham	Sep 2017	
International Conference on Social Informatics 2017	Oxford	Sep 2017	Speaker

continued.....

7. Appendices

.....*continuation*

Name	Location	Date	Role
Gikii 2017	Winchester	Sep 2017	Speaker
AI and Policy roundtable, International Conference of Data Protection and Privacy Commissioners	Hong Kong, HK	Sep 2017	Speaker
Public Sector Machine Learning Seminar, HKUST	Hong Kong, HK	Sep 2017	Speaker
BSI-NESTA Machine Learning Standards Workshop	London	Oct 2017	
A Hippocratic Oath for Data Science? DataKind	London	Oct 2017	
NetGain Algorithmic Governance Workshop	London	Oct 2017	
Mozilla Festival 2017	London	Oct 2017	Speaker
Ethical Auditing for Accountable Automated Decision-Making Workshop, University of Oxford	Oxford	Oct 2017	Speaker
Brussels Privacy Symposium	Brussels, BE	Nov 2017	Speaker
The Ethics of Coding, ICA	London	Nov 2017	Speaker
adigital Outthink 2017	Madrid, ES	Nov 2017	Speaker
UCL Machine Learning Academy	London	Nov 2017	Speaker
New Zealand AI and Law Workshop, University of Oxford	Oxford	Nov 2017	Speaker
Verification and Validation of Ethical Systems Workshop, Cumberland Lodge	Egham	Nov 2017	
EPSRC TIPS Research Community Network	Preston	Dec 2017	Poster
KCL Digital Government Workshop	London	Dec 2017	Speaker
Computers, Privacy and Data Protection 2018 (CPDP)	Brussels, BE	Jan 2018	
PLSC-EU 2018	Brussels, BE	Jan 2018	Speaker
Privacy Camp	Brussels, BE	Jan 2018	Speaker
Social Science Foo Camp, Facebook	Menlo Park, US	Jan 2018	Facilitator
Clarity e-Government Conference	Skelleftea, SE	Feb 2018	Speaker
Civil Service Fast Stream Conference	London	Feb 2018	Speaker
AI and Consent Workshop, Loughborough University	London	Feb 2018	Speaker
Institute of Civil Engineers Conference: What is the City without the People?	London	Feb 2018	Speaker
FAT* 2018	New York	Feb 2018	
GDPR and Beyond, Alan Turing Institute	London	Mar 2018	Speaker
Living in the IoT, IET	London	Mar 2018	Speaker
BILETA, University of Aberdeen	Aberdeen	Apr 2018	Speaker
ACM CHI 2018	Montréal, CA	Apr 2018	Speaker
CHI-GDPR 2018	Montréal, CA	Apr 2018	Speaker
Sensemaking Workshop, CHI	Montréal, CA	Apr 2018	Speaker

continued.....

7.2. Workshops and Conferences Attended

.....continuation

Name	Location	Date	Role
Towards Accountable Systems, Schloss Dagstuhl	Saarbrücken, DE	Apr 2018	Speaker
Privacy Enhancing Technologies Workshop, the Royal Society	London	May 2018	
Defining AI Workshop, University of Amsterdam	Amsterdam, NL	Jun 2018	
CognitionX Conference 2018	London	Jun 2018	Speaker
AI, GDPR and the Law Roundtable, EUI	Florence, IT	Jun 2018	Speaker
AI Now Machine Learning, Inequality, and the Challenge of Bias Roundtable	Berlin, DE	Jul 2018	Speaker
ECPR Standing Group on Regulatory Governance	Lausanne, CH	Jul 2018	Speaker
Supporting Algorithm Accountability using Provenance, Provenance Week, KCL	London	Jul 2018	Speaker
Algorithms in the Justice System, the Law Society	London	Jul 2018	Speaker
Gikii 2018	Vienna, AT	Sep 2018	Speaker
MCCRC Symposium: Compliance by Design, University of Cambridge	Cambridge	Sep 2018	
Amsterdam Privacy Conference	Amsterdam, NL	Oct 2018	Speaker
Roundtable on Smartphone Ecosystems	Amsterdam, NL	Oct 2018	
AI and Government, Government of Scotland	Edinburgh	Oct 2018	Speaker
Data Rights Hackday, Open Rights Group	London	Oct 2018	
Algorithmic Discrimination Workshop, KU Leuven	Leuven, BE	Oct 2018	
Partnership on AI All Partners Meeting	San Francisco, US	Nov 2018	Speaker
The Law of Everything? Expansive Notions of Data Protection Law, Tilburg University	Tilburg, NL	Nov 2018	Speaker
Algorithms and Society Workshop, VUB	Brussels, BE	Dec 2018	Speaker
Law, Policy and the Internet Launch, IALS	London	Dec 2018	Speaker
Cognitas Ergo Sum: Ten Years of Profiling the European Citizen Launch, VUB	Brussels, BE	Dec 2018	Speaker

Bibliography

- Abadi M and others, 'On the Protection of Private Information in Machine Learning Systems: Two Recent Approaches' in *Proceedings of the 30th IEEE Computer Security Foundations Symposium, August 21-25, 2017, Santa Barbara, CA, USA* (2017).
- Abbasi A and Chen H, 'Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace' (2008) 26(2) *ACM Trans. Inf. Syst.* 7:1 DOI: 10/dq78jg.
- Ackerman MS, Cranor LF and Reagle J, 'Privacy in e-commerce: examining user scenarios and privacy preferences' (1999) DOI: 10/bdgc4r.
- Adepeju M, Rosser G and Cheng T, 'Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions – a crime case study' (2016) 30(11) *International Journal of Geographical Information Science* 2133 DOI: 10.1080/13658816.2016.1159684.
- Administrative Data Taskforce, *The UK Administrative Data Research Network: Improving access for research and policy* (Economic and Social Research Council 2012) (<http://www.esrc.ac.uk/files/research/administrative-data-taskforce-adt/improving-access-for-research-and-policy/>).
- Afroz S, Islam AC, Stolerman A, Greenstadt R and McCoy D, 'Doppelgänger Finder: Taking Stylometry to the Underground' in *2014 IEEE Symposium on Security and Privacy* (IEEE 2014) DOI: 10/cwnz.
- Afroz S, Islam AC, Stolerman A, Greenstadt R and McCoy D, 'Doppelgänger finder: Taking stylometry to the underground' in *2014 IEEE Symposium on Security and Privacy (SP)* (2014) DOI: 10/cwnz.
- Agrawal R and Srikant R, 'Privacy-preserving Data Mining' in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (ACM 2000).
- AI Now Institute, *The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term* (2016) (<https://artificialintelligencenow.com/>).
- Alciné J (*Twitter* [[@jackyalcine](https://twitter.com/jackyalcine)], 28th June 2015) (<https://perma.cc/E6UT-K8GL>).
- Aldhous P, 'The Digital Search for Victims of Child Pornography' (2011) 210(2807) *New Scientist* 23 DOI: 10/b3brnn.

Bibliography

- Alegre F, Soldi G, Evans N, Fauve B and Liu J, 'Evasion and obfuscation in speaker recognition surveillance and forensics' in *2nd International Workshop on Biometrics and Forensics* (IEEE 2014) DOI: 10/cwnx.
- Alkire S, *Valuing freedoms: Sen's capability approach and poverty reduction* (Oxford University Press 2005).
- Altieri NA, Pisoni DB and Townsend JT, 'Some Normative Data on Lip-Reading Skills' (2011) 130(1) *The Journal of the Acoustical Society of America* DOI: 10.1121/1.3593376.
- Ananny M, 'Toward an ethics of algorithms: Convening, observation, probability, and timeliness' (2016) 4(1) *Science, Technology & Human Values* 93 DOI: 10/gddv77.
- Andrews L, 'Public Administration, Public Leadership and the Construction of Public Value in the Age of the Algorithm and 'Big Data'' *Public Admin.* DOI: 10/gd25br.
- Andrews R, Diederich J and Tickle AB, 'Survey and critique of techniques for extracting rules from trained artificial neural networks' (1995) 8(6) *Knowledge-Based Systems* 373.
- Andriotis A and Ensign RL, 'U.S. Uses Race Test to Decide Who to Pay in Ally Auto-Loan Pact' (*Wall Street Journal*, 29th October 2015) (<https://perma.cc/NBH8-ERDS>).
- Angwin J (*Twitter* [[@JuliaAngwin](https://twitter.com/JuliaAngwin)], 23rd May 2016) (<https://perma.cc/6QLB-KRFM>) accessed 27th November 2017.
- Angwin J, Larson J, Mattu S and Kirchner L, 'Machine bias' (*ProPublica*, 23rd May 2016) (<http://perma.cc/L4M4-TJQT>).
- Angwin J and Parris Jr T, 'Facebook Lets Advertisers Exclude Users by Race' (*ProPublica*, 28th October 2016) (<https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race>).
- Angwin J, Tobin A and Varner M, 'Facebook (Still) Letting Housing Advertisers Exclude... — ProPublica' (*ProPublica*, 21st November 2017) (<https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin>) accessed 1st October 2018.
- Anonymous, 'Wells Fargo yanks "Community Calculator" service after ACORN lawsuit' [2000] *Credit Union Times* (<https://perma.cc/XG79-9P74>).
- 'The Challenges and Threats of Automated Lip Reading' (*MIT Technology Review*, 11th September 2014) (<https://perma.cc/3MEX-FUL9>).
- Apple Inc, *iOS Security: iOS 10* (Apple Inc 2017) (<https://perma.cc/8EQE-TFW5>).
- 'Privacy Policy' (*apple.com*, 19th November 2017) (<http://perma.cc/3DC2-M7Z5>).
- Ardissono L, Goy A, Petrone G, Segnan M and Torasso P, 'Intrigue: Personalized recommendation of tourist attractions for desktop and hand held devices' (2003) 17(8-9) *Applied Artificial Intelligence* 687.

-
- Ariely D and Berns GS, 'Neuromarketing: The Hope and Hype of Neuroimaging in Business' [2010] (284) *Nature Reviews Neuroscience* DOI: 10/cvjds8.
- Article 29 Data Protection Working Party, *Opinion 4/2007 on the concept of personal data (WP 136)* (2007).
- *The Future of Privacy: Joint contribution to the Consultation of the European Commission on the legal framework for the fundamental right to protection of personal data (02356/09/EN WP 168)* (2009).
 - *Opinion 1/2010 on the Concepts of 'Controller' and 'Processor.'* (WP 169) (2010).
 - *Opinion 13/2011 on Geolocation services on smart mobile devices' (WP 185)* (2011).
 - *Opinion 03/2013 on purpose limitation (wp203)* (2013).
 - *Statement on the role of a risk-based approach in data protection legal frameworks* (2014).
 - *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 (WP 248 rev.01)* (2017).
 - *Opinion 2/2017 on data processing at work (WP 249)* (2017).
 - *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, wp251rev.01* (2018).
 - *Guidelines on Consent under Regulation 2016/679 (wp259rev.01)* (2018).
 - *Guidelines on Transparency under Regulation 2016/679 (wp260rev.01)* (2018).
- Assael YM, Shillingford B, Whiteson S and Freitas N de, 'LipNet: End-to-End Sentence Level Lipreading' (*arXiv preprint*, 2016) (<https://arxiv.org/abs/1611.01599>).
- Association for Computing Machinery, *ACM Code of Ethics and Professional Conduct* (ACM 2018) (<https://perma.cc/2UY9-U8YK>).
- Association of Teachers of Lipreading to Adults, 'Lipreading Interpreting Policy' (ATLA) (<https://atlalipreading.org.uk/about-us/lipreading-interpreting-policy/>) accessed 29th December 2017.
- Ateniese G, Mancini LV, Spognardi A, Villani A, Vitali D and Felici G, 'Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers' (2015) 10(3) *IJSN* 137.
- Ausloos J, 'The Right to Erasure: Safeguard for Informational Self-Determination in a Digital Society?' (Doctoral dissertation, KU Leuven 2018).
- Ausloos J and Dewitte P, 'Shattering one-way mirrors—data subject access rights in practice' (2018) 8(1) *International Data Privacy Law* 4 DOI: 10/cwcf.
- Autoriteit de concurrentie and Bundeskartellamt, *Big Data and Competition* (2016).
- Autoriteit Persoonsgegevens, 'Bluetrace beëindigt overtredingen wifi-tracking na optreden AP' [2017] Autoriteit Persoonsgegevens (Dutch Data Protection Authority) Website (<https://perma.cc/D78Q-WDBJ>).

Bibliography

- Azavea, *HunchLab: Under the Hood* (Author 2015) (<http://cdn.azavea.com/pdfs/hunchlab/HunchLab-Under-the-Hood.pdf>).
- Bambauer J, Muralidhar K and Sarathy R, 'Fool's Gold: An Illustrated Critique of Differential Privacy' (2013) 16 *Vand. J. Ent. & Tech. L.* 701.
- Banisar D, *The right to information and privacy: balancing rights and managing conflicts* (World Bank 2011).
- Bardach E, *A practical guide for policy analysis* (SAGE 2012).
- Barker MCJ, Cunningham S and Shao X, 'An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition' (2006) 120 *The Journal of the Acoustical Society of America* DOI: 10/c7mkv9.
- Barocas S, Hardt M and Narayanan A, *Fairness and Machine Learning* (fairmlbookorg 2018) (<http://www.fairmlbook.org>).
- Barocas S and Nissenbaum H, 'Big Data's End Run around Anonymity and Consent' in *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (Cambridge University Press 2014) DOI: 10/cxvb.
- Barocas S and Selbst AD, 'Big Data's Disparate Impact' (2016) 104 *California Law Review* 671 DOI: 10/gfgq9w.
- Barry A and Born G, 'Interdisciplinarity: Reconfigurations of the social and natural sciences' in *Interdisciplinarity* (Routledge 2013).
- Barua D, 'A time to remember, a time to forget: User controlled, Scalable, Life long user modelling' (Doctoral dissertation, The University of Sydney 2016).
- Baskerville RL and Wood-Harper AT, 'A critical perspective on action research as a method for information systems research' (1996) 11(3) *Journal of Information Technology* 235 DOI: 10/b3r58v.
- Bateson G, *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology* (Chandler Pub Co 1972).
- BBC News, 'SPFL facial recognition cash blow from Scottish Government' (*BBC News*, 25th February 2015) (<http://www.bbc.co.uk/sport/football/35664117>).
- 'AI That Lip-Reads "Better than Humans"' (*BBC News*, 8th November 2016) (<http://www.bbc.co.uk/news/technology-3791113>).
- 'Kent slavery raids 'uncover 21 victims'' (*BBC News*, 6th December 2016) (<https://perma.cc/AM4S-RMHR>).
- Bear HL, Cox SJ and Harvey RW, 'Speaker-independent machine lip-reading with speaker-dependent viseme classifiers' in *The 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing (FAAVSP-2015), Vienna, Austria, September 11-13, 2015* (2015).

-
- Bear HL and Harvey R, 'Phoneme-to-Viseme Mappings: The Good, the Bad, and the Ugly' (2017) 95 *Speech Communication* 40 DOI: 10/gctgp6.
- Bear HL, Harvey R, Theobald B.-J and Lan Y, 'Resolution Limits on Visual Speech Recognition' in *2014 IEEE International Conference on Image Processing (ICIP)* (IEEE 2014) DOI: 10/gfrfdd.
- Bear HL and Taylor S, 'Visual Speech Recognition: Aligning Terminologies for Better Understanding' in *Proceedings of British Machine Vision Conference. 28th British Machine Vision Conference. London, UK, 4-7 September 2017, London, September 4-7 2017* (BMVA Press 2017).
- Behrens JT, 'Principles and procedures of exploratory data analysis' (1997) 2(2) *Psychol. Methods* 131.
- Bellegarda JR and Silverman KEA, 'Fast, language-independent method for user authentication by voice' (*US Patent no 9218809*, 2015) (<https://patents.google.com/patent/US9218809B2/>).
- Bench-Capon T, 'Neural Networks and Open Texture' in *Proceedings of the 4th International Conference on Artificial Intelligence and Law (ICAIL '93, ACM 1993)* DOI: 10/fw933t.
- Bench-Capon T and Sergot M, 'Towards a rule-based representation of open texture in law' in C Walter (ed), *Computer power and legal language* (Quorum Books 1988).
- Bench-Capon T and others, 'A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law' (2012) 20(3) *Artificial Intelligence and Law* 215 DOI: 10/gc7mhr.
- Berendt B and Preibusch S, 'Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence' (2014) 22(2) *Artif. Intell. & Law* 175 DOI: 10/gddxpd.
- Berk R, Heidari H, Jabbari S, Kearns M and Roth A, 'Fairness in Criminal Justice Risk Assessments: The State of the Art' [2018] *Sociological Methods & Research* DOI: 10/gfgt87.
- Bernal N, 'Britain's data commissioner launches investigation into UK use of facial recognition' (*The Telegraph*, 3rd December 2018) (<https://perma.cc/AS6Q-4UPX>).
- Besacier L, Barnard E, Karpov A and Schultz T, 'Automatic Speech Recognition for Under-Resourced Languages: A Survey' (2014) 56 *Speech Communication* 85 DOI: 10/gfpgrw.
- Bevan G and Hood C, 'What's measured is what matters: Targets and gaming in the English public health care system' (2006) 84(3) *Public Admin.* 517 DOI: 10/cww324.

- Bietti E and Binns R, 'Acquisitions in the Third Party Tracking Industry: Competition and Data Protection Aspects' [2018] Preprint available on SSRN (<https://papers.ssrn.com/abstract=3269473>).
- Binns R, 'Data protection impact assessments: A meta-regulatory approach' (2017) 7(1) *International Data Privacy Law* 22 DOI: 10/cvct.
- 'Fairness in Machine Learning: Lessons from Political Philosophy', in *Conference on Fairness, Accountability and Transparency (FAT* 2018)* (PMLR 2018) vol 81.
- Binns R, Lyngs U, Van Kleek M, Zhao J, Libert T and Shadbolt N, 'Third Party Tracking in the Mobile Ecosystem' in *Proceedings of the 10th ACM Conference on Web Science* (ACM 2018) DOI: 10/cwdk.
- Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J and Shadbolt N, 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'18)* (ACM 2018) DOI: 10/cvcp.
- Binns R, Veale M, Van Kleek M and Shadbolt N, 'Like trainer, like bot? Inheritance of bias in algorithmic content moderation' in GL Ciampaglia, A Mashhadi and T Yasseri (eds), *Social Informatics: 9th International Conference, SocInfo 2017, Proceedings, Part II* (Springer 2017) DOI: 10/cvc2.
- Binns R, Zhao J, Van Kleek M and Shadbolt N, 'Measuring third party tracker power across web and mobile' (2018) 18(4) *ACM Transactions on Internet Technology (TOIT)* 52.
- Bizer C, Heath T and Berners-Lee T, 'Linked data—the story so far' in A Sheth (ed), *Semantic services, interoperability and web applications: Emerging concepts* (IGI Global 2009).
- Blackmon GA, 'Problems at the register: Retail collection of personal information and the data breach' (2014) 65 *Case W. Res. L. Rev.* 861.
- Blank A, Cram F, Dare T, Smith B and Vaithianathan R, *Ethical issues for Maori in predictive risk modelling to identify new-born children who are at high risk of future maltreatment* (, Government of New Zealand 2015) (<https://perma.cc/EWA7-VL4C>).
- Blodgett SL and O'Connor B, 'Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English' in *Presented at FAT/ML 2017, Halifax, Canada* (2017) (<https://arxiv.org/abs/1707.00061>).
- Bogen M and Rieke A, *Help Wanted—An Exploration of Hiring Algorithms, Equity and Bias* (, Upturn 2018).
- Bolukbasi T, Chang K, Zou J, Saligrama V and Kalai A, 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings' [2016] 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

-
- Bolukbasi T, Chang K, Zou J, Saligrama V and Kalai AT, 'Quantifying and Reducing Stereotypes in Word Embeddings' [2016] arXiv preprint.
- Bosch BFE and Eijk NANM van, 'Wifi-tracking in de winkel(straat): inbreuk op de privacy?' (2016) 19(251) Privacy & Informatie.
- Bostrom N, *Superintelligence: Paths, dangers, strategies* (Oxford University Press 2014).
- Bovens M, 'Analysing and assessing accountability: A conceptual framework' (2007) 13(4) European L.J. 447 DOI: 10/b4hmbf.
- 'Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism' (2010) 33(5) West European Politics 946 DOI: 10/frq37t.
- Bowker G and Star SL, *Sorting things out: Classification and its consequences* (The MIT Press 1999).
- boyd d, 'Undoing the neutrality of Big Data' (2016) 16 Florida Law Review Forum 226.
- boyd d and Crawford K, 'Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon' (2012) 15(5) Information, Communication & Society 662 DOI: 10/7vq.
- Bradshaw T, 'Apple plays catch-up with iMessage emojis' [2016] Financial Times (<https://perma.cc/LT9T-6FV9>).
- Brandeis L, *Other People's Money, and How Bankers Use it* (National Home Library Foundation 1933).
- Brandom R, 'Self-Driving Cars Are Headed toward an AI Roadblock' (*The Verge*, 3rd July 2018) (<https://www.theverge.com/2018/7/3/17530232/self-driving-ai-winter-full-autonomy-waymo-tesla-uber>).
- Brauneis R and Goodman EP, 'Algorithmic Transparency for the Smart City' (2018) 20 Yale J. Law. & Tech. 103 DOI: 10/cncv.
- Breiman L, 'Random Forests' (2001) 45(1) Machine Learning 5 DOI: 10/d8zjwq.
- Breiman L and others, 'Statistical modeling: The two cultures' (2001) 16(3) Statistical Science 199 DOI: 10/bd86gq.
- Bridge M, 'Siri users are denied access to their data' (*The Times*).
- Brooks H, 'Technology, evolution, and purpose' (1980) 109(1) Daedalus 65.
- Brown I and Marsden CT, *Regulating Code* (MIT Press 2013).
- Brunton F and Nissenbaum H, *Obfuscation: A User's Guide for Privacy and Protest* (The MIT Press 2015).
- Buffat A, 'Street-level bureaucracy and e-government' (2015) 17(1) Public Management Review 149 DOI: 10/gfgrbf.
- Bullington J, "'Affective" Computing and Emotion Recognition Systems: The Future of Biometric Surveillance?' in *InfoSecCS'05* (ACM 2005).
- Bundeskartellamt, *Big Data und Wettbewerb* (2017).

- Bundesministerium des Innern Für Bau und Heimat, *Datenethikkommission* (Government of Germany 2018) (<https://perma.cc/7RC6-J8SD>) accessed 6th October 2018.
- Buolamwini J and Gebru T, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' in *Conference on Fairness, Accountability and Transparency* (2018) (<http://proceedings.mlr.press/v81/buolamwini18a.html>) accessed 1st October 2018.
- Burrell J, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2016) 3(1) *Big Data & Society* DOI: 10.1177/2053951715622512.
- Burton J and van den Broek D, 'Accountable and Countable: Information Management Systems and the Bureaucratization of Social Work' (2009) 39(7) *The British Journal of Social Work* 1326 DOI: 10/bd5whm.
- Busvine D, 'Mozilla co-founder's Brave files adtech complaint against Google' (*Reuters*, 12th September 2018) (<https://perma.cc/FRC8-6L8W>).
- Bygrave LA, 'Automated Profiling: Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling' (2001) 17(1) *Comput. Law & Secur. Rev.* 17 DOI: 10.1016/S0267-3649(01)00104-2.
- 'Data Protection by Design and by Default : Deciphering the EU's Legislative Requirements' (2017) 1(2) *Oslo Law Review* 105.
 - 'Legal Scholarship on Data Protection: Future Challenges and Directions', in C de Terwangne, E Degrave, S Dusollier and R Queck (eds), *Liber amicorum Yves Poulet/Esays in honour of Yves Poulet* (Bruylant 2017) (<https://ssrn.com/abstract=3076747>).
- Cadwalladr C, 'Revealed: how US billionaire helped to back Brexit' (*The Observer*, 25th February 2017) (<https://perma.cc/PJ7A-RMN6>).
- 'Parliament seizes cache of Facebook internal papers' (*The Observer*, 24th November 2018) (<https://perma.cc/T7TH-U8AF>).
- Cadwalladr C and Graham-Harrison E, 'Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach' (*The Observer*, 17th March 2018) (<https://perma.cc/HFC8-WWDT>).
- Calandrino JA, Kilzer A, Narayanan A, Felten EW and Shmatikov V, "'You Might Also Like.'" Privacy Risks of Collaborative Filtering' in *IEEE Symposium on Security and Privacy (SP)* (2011) DOI: 10/bfnjg3.
- Caliskan A, Bryson JJ and Narayanan A, 'Semantics derived automatically from language corpora contain human-like biases' (2017) 356(6334) *Science* 183 DOI: 10/f93cpf.
- Campolo A, Sanfilippo M, Whittaker M and Crawford K, *The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term* (AI Now Institute 2017) (<https://perma.cc/G9AX-XQFN>).

-
- Cao Y and Yang J, 'Towards Making Systems Forget with Machine Unlearning' in *2015 IEEE Symposium on Security and Privacy* (2015) DOI: 10/cwd7.
- Capobiano A, Gonzaga P and Nyeső A, *DAF/COMP(2017)4: Algorithms and collusion - Background note by the Secretariat* (Organisation for Economic Co-operation and Development (OECD) 2017) ([https://one.oecd.org/document/DAF/COMP\(2017\)4/en/pdf](https://one.oecd.org/document/DAF/COMP(2017)4/en/pdf)).
- Carlini N, Liu C, Kos J, Erlingsson Ú and Song D, 'The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets' [2018] arXiv preprint (<https://arxiv.org/abs/1802.08232>).
- Cartwright N and Hardie J, *Evidence-Based Policy: A practical guide to doing it better* (Oxford University Press 2012).
- Cate FH, Cullen P and Mayer-Schönberger V, *Data Protection Principles for the 21st Century: Revising the 1980 OECD Guidelines* (, Oxford Internet Institute 2014) (<https://perma.cc/G5HL-VUPL>).
- Cate FH and Mayer-Schönberger V, 'Notice and consent in a world of Big Data' (2013) 3(2) International Data Privacy Law 67 DOI: 10/cvcv.
- Cavoukian A, *Privacy by Design: The 7 Foundational Principles* (Information and Privacy Commissioner for Ontario 2010).
- Chabert J-L, Barbin É, Borowczyk J, Guillemot M, Michel-Pajus A, Djebbar A and Martzloff J.-C, *A history of algorithms: From the pebble to the microchip* (Weeks C tr, Springer 1999).
- Chakaravarthy VT, Gupta H, Roy P and Mohania MK, 'Efficient Techniques for Document Sanitization' in *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08, ACM 2008)* DOI: 10/frzwmc.
- Chaudhuri K, Monteleoni C and Sarwate AD, 'Differentially private empirical risk minimization' (2011) 12 J. Mach. Learn. Res. 1069.
- Chen H and others, 'Visualization in law enforcement' in *CHI'05 Extended Abstracts on Human Factors in Computing Systems* (ACM 2005).
- Cheyner AJ, 'Device access using voice authentication' (*US Patent no 9262612*, 2016) (<https://patents.google.com/patent/US9262612B2/>).
- Chouldechova A, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments' (2017) 5(2) Big Data 153 DOI: 10/gdcdqf.
- Chouldechova A, Benavides-Prado D, Fialko O and Vaithianathan R, 'A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions' in *Conference on Fairness, Accountability and Transparency (FAT* 2018)* (2018).
- Christie GC, 'An essay on discretion' (1986) 5 Duke Law Journal 747.
- Chung JS, Senior A, Vinyals O and Zisserman A, 'Lip Reading Sentences in the Wild' (*arXiv preprint*, 2016) (<http://arxiv.org/abs/1611.05358>).

- Citron DK, 'Technological due process' (2008) 85 Washington University Law Review 1249.
- Clancey WJ, 'The epistemology of a rule-based expert system—a framework for explanation' (1983) 20(3) Artificial intelligence 215.
- Clifford D and Ausloos J, 'Technobabble and Technobullsh*t—what the hell is everyone on about?' in *Presented at Gikii 2018, September 15 2017, Winchester, UK* (2017).
- Cobbe J, 'Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making' [2018] Available on SSRN DOI: 10 / gd25bq.
- Coglianese C and Lehr D, 'Regulating by Robot: Administrative Decision Making in the Machine-Learning Era' (2016) 105 Geo. LJ 1147 (<https://ssrn.com/abstract=2928293>).
- Cohen JE, 'Examined lives: Informational privacy and the subject as object' (2000) 52 Stanford Law Rev. 1373.
- Cohn JF and De La Torre F, 'Automated Face Analysis for Affective Computing' in R Calvo, S D'Mello, J Gratch and A Kappas (eds), *The Oxford Handbook of Affective Computing* (Oxford University Press 2015).
- College bescherming persoonsgegevens, *Wifi-tracking van mobiele apparaten in en rond winkels door Bluetrace (Rapport z2014-00944)* (Autoriteit Persoonsgegevens (Dutch Data Protection Authority) October 2015) (<https://perma.cc/6BKJ-JGPY>).
- Commission, 'Amended proposal for a Council Directive on the protection of individuals with regard to the processing of personal data and on the free movement of such data' COM(92) 422 final—SYN 297.
- Commission, 'Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market' COM(2016) 593 final.
- Commission, 'Proposal for a Directive of the European Parliament and of the Council on representative actions for the protection of the collective interests of consumers, and repealing Directive 2009/22/EC' COM (2018) 184 final.
- Commission, 'Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications)' COM(2018) 640 final.
- Commission, 'Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online' COM(2018) 640 final.
- Commission, 'Proposal for a Regulation of the European Parliament and of the Council on promoting fairness and transparency for business users of online intermediation services' COM(2018) 238 final.

-
- Commission nationale de l'informatique et des libertés (CNIL), *8e Rapport au président de la République et au Parlement, 1987* (La Documentation Française 1988) (<https://perma.cc/2NCW-R5Q3>).
- Committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT), *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework (MSI-AUT(2018)05)* (, Council of Europe 2018).
- Committee of experts on internet intermediaries (MSI-NET), *Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications (MSI-NET(2016)06 rev3 FINAL)* (, Council of Europe 2017) (<https://perma.cc/GS4B-ZYHA>).
- Committee on Civil Liberties, Justice and Home Affairs, *Report on the proposal for a regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications)* (European Parliament 2017).
- 'Common Statement by the Contact Group of the Data Protection Authorities of The Netherlands, France, Spain, Hamburg and Belgium' (CNIL, May 2017) (<https://www.cnil.fr/fr/node/23602>).
- Conigliaro D, Setti F, Bassetti C, Ferrario R and Cristani M, 'ATTENTO: ATTENTION Observed for Automated Spectator Crowd Analysis' in AA Salah, H Hung, O Aran and H Gunes (eds), *Human Behavior Understanding* (Lecture Notes in Computer Science, Springer International Publishing 2013).
- Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, *Artificial Intelligence and Data Protection: Challenges and Possible Remedies (T-PD(2018)09Rev)* (, Council of Europe 2018).
- Cooke N, 'Varieties of knowledge elicitation techniques' (1994) 41(6) *International Journal of Human-Computer Studies* 801 DOI: 10.1006/ijhc.1994.1083.
- Cormack A, 'Is the Subject Access Right Now Too Great a Threat to Privacy' (2016) 2 *Eur. Data Prot. L. Rev.* 15.
- Cormen TH, Leiserson CE, Rivest RL and Stein C, *Introduction to algorithms* (MIT Press 2009).
- Cox SJ, Harvey R, Lan Y, Newman J and Theobald B.-J, 'The Challenge of Multispeaker Lip-Reading' in *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)* (2008) (<https://perma.cc/LQ9X-G6F3>).

- Crabtree A, Lodge T, Colley J, Greenhalgh C, Mortier R and Haddadi H, 'Enabling the new economic actor: data protection, the digital economy, and the Databox' (2016) 20(6) *Personal and Ubiquitous Computing* 947 DOI: 10/f9b3hp.
- Crawford K, 'The Problem with Bias' [2017] Keynote given at NIPS 2017.
- Crawford K and Schultz J, 'Big Data and due process: Toward a framework to redress predictive privacy harms' (2014) 55 *Boston College Law Review* 93.
- Crenshaw K, 'Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics' [1989] *U. Chi. Legal F.* 139.
- Cristianini N and Shawe-Taylor J, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge University Press 2000).
- Crowcroft J, Madhavapeddy A, Schwarzkopf M, Hong T and Mortier R, 'Unclouded Vision' in MK Aguilera, H Yu, NH Vaidya, V Srinivasan and RR Choudhury (eds), *Distributed Computing and Networking* (Springer 2011) DOI: 10/bj9n6t.
- Cuccaro-Alamin S, Foust R, Vaithianathan R and Putnam-Hornstein E, 'Risk Assessment and Decision Making in Child Protective Services: Predictive Risk Modeling in Context' (2017) 79 *Children and Youth Services Review* 291 DOI: 10/gc6c6n.
- Dahlström C, Lapuente V and Teorell J, 'The Merit of Meritocratization: Politics, Bureaucracy, and the Institutional Deterrents of Corruption' (2011) 65(3) *Polit. Res. Q.* 656.
- Danaher J, 'The threat of algocracy: Reality, resistance and accommodation' (2016) 29 *Philosophy & Technology* 245 DOI: 10/gddv8k.
- Dasu T and Johnson T, *Exploratory data mining and data cleaning* (John Wiley & Sons 2003).
- Datta A, Datta A, Makagon J, Mulligan DK and Tschantz MC, 'Discrimination in Online Advertising: A Multidisciplinary Inquiry' in *Conference on Fairness, Accountability and Transparency* (2018) (<http://proceedings.mlr.press/v81/datta18a.html>).
- Datta A, Tschantz MC and Datta A, 'Automated Experiments on Ad Privacy Settings' (2015) 2015(1) *Proceedings on Privacy Enhancing Technologies* 92 DOI: 10/gcv7m7.
- Datta A, Sen S and Zick Y, 'Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems' in *2016 IEEE Symposium on Security and Privacy (SP)* (2016).
- Davis A, Rubinstein M, Wadhwa N, Mysore GJ, Durand F and Freeman WT, 'The Visual Microphone: Passive Recovery of Sound from Video' (2014) 33(4) *ACM Trans. Graph.* 79:1 DOI: 10/gddhjm.
- 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).

-
- De Hert P and Gutwirth S, 'Data Protection in the Case Law of Strasbourg and Luxembourg: Constitutionalisation in Action' in S Gutwirth, Y Poullet, P De Hert, C de Terwangne and S Nouwt (eds), *Reinventing Data Protection?* (Springer 2009) DOI: 10/d4n22h.
- de Montjoye Y.-A, Radaelli L, Singh VK and Pentland A, 'Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata' (2015) 347(6221) *Science* 536 DOI: 10/zt7.
- Dehak N, Kenny PJ, Dehak R, Dumouchel P and Ouellet P, 'Front-End Factor Analysis for Speaker Verification' (2011) 19(4) *IEEE Trans. Audio Speech Lang. Processing* 788.
- Delgado R, 'The Imperial Scholar Revisited: How to Marginalise Outsider Writing, Ten Years Later' (1992) 140 *U. Penn. L. Rev.* 1349.
- Dencik L, Hintz A, Redden J and Warne H, *Data Scores as Governance: Investigating uses of citizen scoring in public services* (, Data Justice Lab, Cardiff University 2018).
- Deng L and Li X, 'Machine learning paradigms for speech recognition: An overview' (2013) 21(5) *IEEE Transactions on Audio, Speech, and Language Processing* 1060.
- Deng L and others, 'Recent Advances in Deep Learning for Speech Research at Microsoft' in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013) DOI: 10/gcsgbd.
- Denham E, 'Consent is not the 'silver bullet' for GDPR compliance' [2016] Information Commissioner's Office Blog (<https://iconewsblog.org.uk/2017/08/16/consent-is-not-the-silver-bullet-for-gdpr-compliance/>) accessed.
- Denhardt JV and Denhardt RB, 'The New Public Service Revisited' (2015) 75(5) *Public Adm. Rev.* 664.
- Denhardt RB and Denhardt JV, 'The New Public Service: Serving Rather than Steering' (2000) 60(6) *Public Adm. Rev.* 549.
- Department for Digital, Culture, Media & Sport, *Centre for Data Ethics and Innovation Consultation* (HM Government 2018) (<https://perma.cc/GG22-GA4Q>) accessed 6th October 2018.
- *Data Ethics Framework* (HM Government 2018).
- Desai DR and Kroll JA, 'Trust But Verify: A Guide to Algorithms and the Law' (2018) 31 *Harv. J.L. & Tech.* 1.
- Devet C and Goldberg I, 'The Best of Both Worlds: Combining Information-Theoretic and Computational PIR for Communication Efficiency' in *Privacy Enhancing Technologies* (Springer 2014).
- Diakopoulos N, 'Algorithmic accountability' (2014) 3(3) *Digital Journalism* 398 DOI: 10/gc5t4g.

- Diaz C, Tene O and Gürses S, 'Hero or villain: The data controller in privacy law and technologies' (2013) 74 Ohio St. LJ 923.
- Dietrich W, Mendoza C and Brennan T, *COMPAS risk scales: Demonstrating accuracy equity and predictive parity* (Northpointe 2016).
- Dietvorst BJ, Simmons JP and Massey C, 'Algorithm aversion: People erroneously avoid algorithms after seeing them err' (2015) 144(1) Journal of Experimental Psychology: General 114 DOI: 10/f6xqfw.
- Differential Privacy Team, Apple, 'Learning with Privacy at Scale' (2017) 1(8) Apple Machine Learning Journal (<https://perma.cc/T2RM-B27X>).
- Dijkstra JJ, 'User agreement with incorrect expert system advice' (1999) 18(6) Behaviour & Information Technology 399 DOI: 10/fsnqm9.
- DiMaggio PJ and Powell WW, 'The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields' (1983) 48(2) American Sociological Review 147.
- Ding B, Kulkarni J and Yekhanin S, 'Collecting Telemetry Data Privately' in I Guyon, U Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett (eds), *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc 2017).
- Doctorow C, 'Data protection in the EU: the certainty of uncertainty' [2013] The Guardian (<https://perma.cc/DFY4-9SNC>).
- Domingos P, 'A few useful things to know about machine learning' (2012) 55(10) Commun. ACM 78 DOI: 10/cgc9.
- Doshi-Velez F and others, 'Accountability of AI under the law: The role of explanation' [2017] arXiv preprint (<https://arxiv.org/abs/1711.01134>).
- doteveryone, *People, Power and Technology: The 2018 Digital Attitudes Report* (doteveryone 2018) (<https://perma.cc/WT2C-SJ75>).
- Douglas HE, 'The Moral Responsibilities of Scientists (Tensions between Autonomy and Responsibility)' (2003) 40(1) American Philosophical Quarterly 59.
- Dourish P, 'Accounting for system behaviour: Representation, reflection and resourceful action' [1997] Computers and Design in Context 145.
- 'Algorithms and their others: Algorithmic culture in context' (2016) 3(2) Big Data & Society DOI: 10/gcdx9q.
- Downs A, 'A Realistic Look at the Final Payoffs from Urban Data Systems' (1967) 27(3) Public Adm. Rev. 204 DOI: 10.2307/973283.
- Doyle D, Tsybmal A and Cunningham P, *A Review of Explanation and Explanation in Case-Based Reasoning* (, Department of Computer Science, Trinity College, Dublin 2003).
- Duch W, 'Coloring black boxes: visualization of neural network decisions' (2003) 3 International Joint Conference on Neural Networks 1735.

-
- Duhigg C, 'How Companies Learn Your Secrets' (*New York Times Magazine*, 16th February 2012) (<https://perma.cc/2E69-JRKW>).
- Dunleavy P and Hood C, 'From old public administration to new public management' (1994) 14(3) *Public Money & Management* 9.
- Dunleavy P, Margetts H, Bastow S and Tinkler J, *Digital Era Governance: IT Corporations, the State and e-Government* (Oxford University Press 2006) DOI: 10/fhj5bz.
- 'New Public Management is Dead – Long Live Digital-Era Governance' (2006) 16(3) *J. Public Adm. Res. Theory* 467 DOI: 10.1093/jopart/mui057.
- Dwork C, 'Differential Privacy: A Survey of Results' in M Agrawal, D Du, Z Duan and A Li (eds), *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings* (Springer 2008) DOI: 10.1007/978-3-540-79228-4_1.
- Dwork C, Hardt M, Pitassi T, Reingold O and Zemel R, 'Fairness through awareness' in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)* (2012) DOI: 10/fzd3f9.
- Dwork C, McSherry F, Nissim K and Smith A, 'Calibrating Noise to Sensitivity in Private Data Analysis' in S Halevi and T Rabin (eds), *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings* (Springer 2006).
- Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG and Beck HP, 'The role of trust in automation reliance' (2003) 58(6) *International Journal of Human-Computer Studies* 697 DOI: 10/cgvddv.
- Eck D, 'NIPS 2016 & Research at Google' (*Google Research Blog*, 4th December 2016) (<https://research.googleblog.com/2016/12/nips-2016-research-at-google.html>).
- Edwards L, 'Modelling Law Using a Feminist Theoretical Perspective' (1995) 4(1) *Information & Communications Technology Law* 95 DOI: 10/bxwnwf.
- 'Privacy, security and data protection in smart cities: A critical EU law perspective' (2016) 2 *Eur. Data Prot. L. Rev.* 28.
- 'Data protection: Enter the General Data Protection Regulation', in L Edwards (ed), *Law, Policy and the Internet* (Hart 2018).
- Edwards L and Harbinja E, 'Protecting Post-Mortem Privacy: Reconsidering the Privacy Interests of the Deceased in a Digital World' (2013) 32 *Cardozo Arts & Ent L J* 83 DOI: 10/gfhxns.
- Edwards L 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 L. & Tech. Rev.* 18 DOI: 10/gdxthj.

Bibliography

- Edwards L and Veale M, 'Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?' (2018) 16(3) IEEE Security & Privacy 46 DOI: 10/gdz29v.
- Eisman Lott B and Levy J, 'The Influence of Certain Communicator Characteristics on Lip Reading Efficiency' (1960) 51 The Journal of Social Psychology 419 DOI: 10.1080/00224545.1960.9922051.
- Electronic Frontier Foundation, 'Do Not Track' (*eff.org*, 31st December 2018) (<https://perma.cc/CF2M-YFZR>).
- Elish MC, 'Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction' (2019) 5 Engaging Science, Technology, and Society 40 DOI: 10/gf2t99.
- Elliot M, Mackey E, O'Hara K and Tudor C, *The Anonymisation Decision-Making Framework* (UKAN 2016).
- Elliot M and others, 'Functional Anonymisation: Personal Data and the Data Environment' (2018) 34(2) Computer Law & Security Review 204 DOI: 10/gdhs4w.
- Ensign D, Friedler SA, Neville S, Scheidegger C and Venkatasubramanian S, 'Runaway Feedback Loops in Predictive Policing' in *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAT*)* (2018).
- Ephrat A, Halperin T and Peleg S, 'Improved Speech Reconstruction from Silent Video' in *ICCV 2017 Workshop on Computer Vision for Audio-Visual Media* (2017) (<https://arxiv.org/abs/1708.01204>).
- Erlingsson Ú, Pihur V and Korolova A, 'Rappor: Randomized aggregatable privacy-preserving ordinal response' in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (2014) DOI: 10/cwj2.
- Eslami M, Krishna Kumaran SR, Sandvig C and Karahalios K, 'Communicating Algorithmic Process in Online Behavioral Advertising' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM 2018) DOI: 10/cxrf.
- Esvelt KM, Smidler AL, Catteruccia F and Church GM, 'Concerning RNA-Guided Gene Drives for the Alteration of Wild Populations' (2014) 3 eLife DOI: 10/gfphx6.
- European Commission, *A Digital Agenda for Europe (COM(2010)245 final)* (2010).
- Special Eurobarometer 431: "Data Protection" (European Union 2015) DOI: 10.2838/552336.
- European Commission, *Artificial Intelligence—a European perspective* (2018) DOI: 10.2760/11251.
- European Commission, *Automated Decision-Making on the Basis of Personal Data That Has Been Transferred from the EU to Companies Certified under the EU-U.S. Privacy Shield: Fact-Finding and Assessment of Safeguards Provided by U.S. Law* (2018).

-
- European Data Protection Supervisor, *Privacy and Competitiveness in the Age of Big Data: The Interplay between Data Protection, Competition Law and Consumer Protection in the Digital Economy* (EDPS 2014).
- *Meeting the challenges of big data, Opinion 7/2015* (EDPS 2015).
 - *Opinion 6/2017: EDPS Opinion on the Proposal for a Regulation on Privacy and Electronic Communications (ePrivacy Regulation)* (EDPS 2017).
 - *Opinion 8/2018 on the legislative package “A New Deal for Consumers”* (EDPS 2018).
- Facebook, Inc, Letter from Facebook Inc. to Chairman Greg Walden: “House Energy and Commerce Questions for the Record” (June 2018) (<https://perma.cc/K6TM-W2N2>).
- Fahy R, Hoboken J van and Eijk N van, ‘Data Privacy, Transparency and the Data-Driven Transformation of Games to Services’ in *2018 IEEE Games, Entertainment, Media Conference (GEM)* (IEEE 2018) DOI: 10/gfsvz6.
- Farhan and others, ‘Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data’ in *2016 IEEE Wireless Health (WH)* (2016) DOI: 10/cwdh.
- Feige U, Fiat A and Shamir A, ‘Zero-knowledge proofs of identity’ (1988) 1(2) *Journal of Cryptology* 77.
- Feldman M, Friedler SA, Moeller J, Scheidegger C and Venkatasubramanian S, ‘Certifying and removing disparate impact’ in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015) DOI: 10/gfgrbk.
- Fernandes M, Walls L, Munson S, Hullman J and Kay M, ‘Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making’ in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI ’18, ACM 2018) DOI: 10/gfrfcn.
- Fischer F, ‘Alexa, tell me my secrets’ (*Projects by If*, 28th November 2018) (<https://www.projectsbyif.com/blog/alexa-tell-me-my-secrets>).
- Flavio Chierichetti Ravi Kumar SL and Vassilvitskii S, ‘Fair Clustering Through Fairlets’ in *Presented as a talk at the 4th Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT/ML 2017), Halifax, Nova Scotia* (2017).
- Forsythe DE, ‘Using ethnography in the design of an explanation system’ (1995) 8(4) *Expert Systems with Applications* 403 DOI: 10/c924fs.
- ‘New bottles, old wine: Hidden cultural assumptions in a computerized explanation system for migraine sufferers’ (1996) 10(4) *Medical Anthropology Quarterly* 551 DOI: 10/b4hp3p.

- Fraunhofer Institute for Integrated Circuits, 'Face Detection Software SHORE: Fast, Reliable and Real-time Capable' (*Fraunhofer IIS*, 2018) (<https://perma.cc/9PWR-QGK6>) accessed 3rd December 2018.
- 'Fraunhofer IIS Presents an Intelligent Sensor for Anonymous Video Analysis at the MWC in Barcelona' (27th February 2018) (<https://perma.cc/Y3VS-59VZ>) accessed 3rd December 2018.
- Fredrikson M, Jha S and Ristenpart T, 'Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures' in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (2015) DOI: 10/cwadm.
- Fredrikson M, Lantz E, Jha S, Lin S, Page D and Ristenpart T, 'Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing' in *USENIX Security Symposium* (2014).
- Freitas J, Teixeira A, Dias MS and Silva S, *Introduction to Silent Speech Interfaces* (Springer 2017) DOI: 10/gftgix.
- Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP and Roth D, 'A Comparative Study of Fairness-Enhancing Interventions in Machine Learning' [2018] arXiv preprint (<http://arxiv.org/abs/1802.04422>).
- Fung A, Graham M and Weil D, *Full Disclosure: The Perils and Promise of Transparency* (Cambridge University Press 2007).
- Funtowicz SO and Ravetz JR, 'Science for the Post-Normal Age' (1993) 25(7) *Futures* 739 DOI: 10/fqntk9.
- Future of Privacy Forum, *Mobile Location Analytics Code of Conduct* (FPF 2013) (<https://perma.cc/LC4B-FHY5>).
- *About Mobile Location Analytics Technology* (FPF 2016).
- Galetta A and Hert P de, 'Exercising Access Rights in Belgium' in C Norris, P de Hert, X L'Hoiry and A Galetta (eds), *The Unaccountable State of Surveillance: Exercising Access Rights in Europe* (Springer 2017).
- Gama J, Žliobaitė I, Bifet A, Pechenizkiy M and Bouchachia A, 'A survey on concept drift adaptation' (2013) 1(1) *ACM Comput. Surv.* DOI: 10/gd893p.
- Gandy OH, *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage* (Routledge 2009).
- 'Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems' (2010) 12(1) *Ethics and Information Technology* 29 DOI: 10/bzwqrx.
- Gao B, 'Exploratory Visualization Design Towards Online Social Network Privacy and Data Literacy' (PhD, KU Leuven 2015).

-
- Gao B and Berendt B, 'Visual data mining for higher-level patterns: discrimination-aware data mining and beyond' in *Proceedings of the 20th Machine Learning conference of Belgium and The Netherlands (Benelearn 2011)* (2011) (<https://perma.cc/E7R6-LYQK>).
- Gebhart G and Williams J, 'Facebook Doesn't Need To Listen Through Your Microphone To Serve You Creepy Ads' (*Electronic Frontier Foundation (EFF) Deeplinks Blog*, 13th April 2018) (<https://perma.cc/QU7B-LA98>).
- Gellert R, de Vries K, de Hert P and Gutwirth S, 'A Comparative Analysis of Anti-Discrimination and Data Protection Legislations' in B Custers, T Calders, B Schermer and T Zarsky (eds), *Discrimination and privacy in the information society* (Springer 2013).
- Gelman A and Hill J, 'Missing-Data Imputation' in *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press 2006) DOI: 10/dv8cb9.
- Georgiadis M, Lane K and Whalley S, *Smart data, smart decisions, smart profits: The retailers' advantage* (McKinsey & Company 2000) (<https://perma.cc/7T5W-WSYB>).
- German Expert Council on Consumer Affairs, *Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren* (2018).
- Ghosh D, 'What You Need to Know About California's New Data Privacy Law' (*Harvard Business Review*, 11th July 2018) (<https://perma.cc/DZ4V-AX4U>).
- Gibbs S, 'Google Alters Search Autocomplete to Remove 'are Jews Evil' Suggestion' (*The Guardian*, 5th December 2016) (<https://perma.cc/P4GF-HTQB>).
- Gibney E, 'AI talent grab sparks excitement and concern' (2016) 532(7600) *Nature* 422 DOI: 10/bfrc.
- 'Google AI Algorithm Masters Ancient Game of Go' (2016) 529(7587) *Nature* 445 DOI: 10/bb5s.
- Goldweber M, Davoli R, Little JC, Riedesel C, Walker H, Cross G and Von Kinsky BR, 'Enhancing the social issues components in our computing curriculum: Computing for the social good' (2011) 2(1) *ACM Inroads* 64.
- Gomez-Barrero M, Rathgeb C, Galbally J, Fierrez J and Busch C, 'Protected Facial Biometric Templates Based on Local Gabor Patterns and Adaptive Bloom Filters' in *Proceedings of the 2014 22nd International Conference on Pattern Recognition (ICPR '14)*, IEEE Computer Society 2014) DOI: 10/gfkscn.
- González Fuster G, *The Emergence of Personal Data Protection as a Fundamental Right of the EU* (Springer 2014) DOI: 10/gfgrb2.
- Google, 'Detecting Faces: Google Cloud Vision API Documentation' (*Google Cloud Platform*, 15th August 2017) (<https://perma.cc/M82U-TAVH>).
- 'Manage Google Voice & Audio Activity' (*Google Search Help*, 2017) (<https://perma.cc/BEJ3-PM3G>).

Bibliography

- Google, *Transparency Report* (<https://perma.cc/8DE4-AXBW>).
- S Goonatilake and S Khebbal (eds), *Intelligent Hybrid Systems* (John Wiley & Sons, Inc 1994).
- Graham DA, 'Not Even Cambridge Analytica Believed its Hype' (*The Atlantic*, 20th March 2018) (<https://perma.cc/7G5B-2PPF>).
- Green L, *The Rationale of Proximate Cause* (Vernon Law Book Company 1927).
- Greenleaf G, 'European' Data Privacy Standards Implemented in Laws Outside Europe' (2018) 149 *Privacy Laws & Business International Report* 21 (<https://ssrn.com/abstract=3096314>).
- Greis M, Hullman J, Correll M, Kay M and Shaer O, 'Designing for Uncertainty in HCI: When Does Uncertainty Help?' in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '17, ACM 2017) DOI: 10/gfn45s.
- Guardans I, 'A New (Sort of) Class Action in Protection of European Consumers' (*Lexology*, 25th April 2018) (<https://perma.cc/X2L2-9J4Y>).
- Guidotti R, Monreale A, Turini F, Pedreschi D and Giannotti F, 'A Survey Of Methods For Explaining Black Box Models' [2018] arXiv preprint (<https://arxiv.org/abs/1802.01933v>).
- Guinchard A, 'The Computer Misuse Act 1990 to Support Vulnerability Research? Proposal for a Defence for Hacking as a Strategy in the Fight against Cybercrime.' (2018) 2(2) *Journal of Information Rights, Policy and Practice* DOI: 10/gfscft.
- Gürses S, 'Can You Engineer Privacy?' (2014) 57(8) *Commun. ACM* 20 DOI: 10/gdxwh5.
- Gürses S, Overdorf R and Balsa E, 'Stirring the POTs: protective optimization technologies' in E Bayamlioglu, I Baraliuc, LAW Janssens and M Hildebrandt (eds), *BEING PROFILED: COGITAS ERGO SUM* (Amsterdam University Press 2018).
- Gürses S and van Hoboken J, 'Privacy after the Agile Turn' in E Selinger, J Polonetsky and O Tene (eds), *The Cambridge Handbook of Consumer Privacy* (Cambridge University Press 2018) DOI: 10/gfgq84.
- Hacking I, 'The looping effects of human kinds' in D Premack, D Sperber and AJ Premack (eds), *Causal cognition: A multidisciplinary debate* (Oxford University Press 1995).
- 'Kinds of People: Moving Targets', in P Marshall (ed), *Proceedings of the British Academy, Volume 151* (British Academy 2007) DOI: 10/gfgrbm.
- Hajian S and Domingo-Ferrer J, 'Direct and indirect discrimination prevention methods' in B Custers, T Calders, B Schermer and T Zarsky (eds), *Discrimination and privacy in the information society* (Springer 2012).
- 'A Methodology for Direct and Indirect Discrimination Prevention in Data Mining' (2013) 25(7) *IEEE Trans. Knowl. Data Eng.* 1445 DOI: 10/f4xxs6.

-
- Hajian S, Domingo-Ferrer J and Farràs O, 'Generalization-based privacy preservation and discrimination prevention in data publishing and mining' (2014) 28(5-6) *Data Min. Knowl. Discov.* 1158 DOI: 10/f6fs97.
- Hall W and Pesenti J, *Growing the artificial intelligence industry in the UK* (HM Government 2017) (<https://perma.cc/3E45-MYSM>).
- Hallsworth M, List JA, Metcalfe RD and Vlaev I, 'The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance' (2017) 148 *Journal of Public Economics* 14 DOI: 10/f96mbc.
- Hamer D, 'Factual Causation' and 'Scope of Liability': What's the Difference?' (2014) 77(2) *The Modern Law Review* 155 DOI: 10/gfv2qt.
- Hamlyn R, Matthews P and Shanahan M, *Science Education Tracker: Young people's awareness and attitudes towards machine learning* (The Wellcome Trust, the Royal Society, and the Department for Business, Energy & Industrial Strategy 2017) DOI: 10/cxnf.
- Hansen M, Jensen M and Rost M, 'Protection Goals for Privacy Engineering' in *2015 IEEE Security and Privacy Workshops* (IEEE 2015) DOI: 10/cwmw.
- Harcourt BE, *Against prediction: Profiling, policing, and punishing in an actuarial age* (University of Chicago Press 2006).
- Hardesty L, 'Milestone for MIT Press's bestseller' (*MIT News*, 10th August 2011) (<http://news.mit.edu/2011/introduction-to-algorithms-500k-0810>).
- Hardt M, Price E and Srebro N, 'Equality of Opportunity in Supervised Learning' in D Lee, Sugiyama, U Luxburg, Guyon and Garnett (eds), *Advances in Neural Information Processing Systems 29* (Curran Associates, Inc 2016).
- Hasan R, Hassan E, Li Y, Caine K, Crandall DJ, Hoyle R and Kapadia A, 'Viewer Experience of Obscuring Scene Elements in Photos to Enhance Privacy' in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18, ACM 2018) DOI: 10/gfkr9s.
- Hassner T, Itcher Y and Kliper-Gross O, 'Violent Flows: Real-Time Detection of Violent Crowd Behavior' in *2012 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (IEEE 2012) DOI: 10/gfrfc8.
- Haughay AP, 'User profiling for voice input processing' (*US Patent no 9633660*, 2017) (<https://patents.google.com/patent/US9190062B2/>).
- Hayes GR, 'The relationship of action research to human-computer interaction' (2011) 18(3) *ACM Transactions on Computer-Human Interaction* (TOCHI) 15 DOI: 10.1145/1993060.1993065.
- Helberger N, Zuiderveen Borgesius F and Reyna A, 'The Perfect Match? A Closer Look at the Relationship between EU Consumer Law and Data Protection Law' (2017) 54(5) *Common Market Law Review* 1427.

- Henderson T, 'Does the GDPR Help or Hinder Fair Algorithmic Decision-Making?' (LLM, University of Edinburgh 2017) DOI: 10/cx88.
- Hendler J and Berners-Lee T, 'From the Semantic Web to social machines: A research challenge for AI on the World Wide Web' (2010) 174(2) *Artificial Intelligence* 156 DOI: 10/cr4qpw.
- Hernandez J, Hoque M, Drevo W and Picard R, 'Mood Meter: Counting Smiles in the Wild' in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp'12)* (ACM 2012) DOI: 10/gfrfdb.
- Heurix J, Zimmermann P, Neubauer T and Fenz S, 'A taxonomy for privacy enhancing technologies' (2015) 53 *Comput. Secur.* 1 DOI: 10/f74pgf.
- Hildebrandt M, 'Profiling and the rule of law' (2008) 1(1) *Identity in the Information Society* 55 DOI: 10.1007/s12394-008-0003-1.
- 'Who is profiling who? Invisible visibility', in *Reinventing Data Protection?* (Springer 2009) DOI: 10/ft4fnz.
 - 'The Dawn of a Critical Transparency Right for the Profiling Era', in J Bus, M Crompton, M Hildebrandt and G Metakides (eds), *Digital Enlightenment Yearbook* (IOS Press 2012) DOI: 10/cwpm.
 - 'Profile Transparency by Design?: Re-Enabling Double Contingency', in M Hildebrandt and K de Vries (eds), *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology* (Routledge 2013).
 - 'Slaves to Big Data. Or Are We?' (2013) 17 *IDP Revista de Internet Derecho y Política* 27 DOI: 10/gd82jr.
 - *Smart technologies and the End(s) of Law* (Edward Elgar 2015).
- Hildebrandt M and Tielemans L, 'Data protection by design and technology neutral law' (2013) 29(5) *Comput. Law & Secur. Rev.* 509.
- Hill R, 'Algorithms, Henry VIII powers, dodgy 1-man-firms: Reg strokes claw over Data Protection Bill' (*The Register*, 30th October 2017) <<https://perma.cc/A6U2-N9VU>>.
- 'Chap asks Facebook for data on his web activity, Facebook says no, now watchdog's on the case' (*The Register*, 24th August 2018) <<https://perma.cc/AT5V-VSEU>>.
- Hillman AJ and Keim GD, 'Shareholder value, stakeholder management, and social issues: what's the bottom line?' (2001) 22(2) *Strategic Management Journal* 125.
- Hirschman AO, *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States* (Harvard University Press 1970).
- Hodson H, 'Google's DeepMind AI can lip-read TV shows better than a pro' (*New Scientist*, 21st November 2016) <<https://perma.cc/63HM-G2D8>>.

-
- Hoens T, Polikar R and Chawla NV, 'Learning from streaming data with concept drift and imbalance: An overview' (2012) 1(1) *Progress in Artificial Intelligence* 89 DOI: 10/fx4rz5.
- Hoepman J.-H, 'Privacy Design Strategies' in *ICT Systems Security and Privacy Protection* (Springer 2014).
- Hoffman RR, 'Human Factors Contributions to Knowledge Elicitation' (2008) 50(3) *Human Factors* 481 DOI: 10/ccd4v3.
- Hoffman RR, Crandall B and Shadbolt N, 'Use of the critical decision method to elicit expert knowledge: A case study in the methodology of cognitive task analysis' (1998) 40(2) *Human Factors* 254 DOI: 10.1518/001872098779480442.
- Holstein K, Wortman Vaughan J, Daumé H, Dudík M and Wallach H, 'From Audit to Action: Design Needs for Fairness Monitoring and Decision Support in Machine Learning Product Teams' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM 2019) DOI: 10.1145/3290605.3300830.
- Hood C, 'A public management for all seasons?' (1991) 69 *Public Admin.* 3 DOI: 10/bdwbfj.
- *The art of the state: Culture, rhetoric, and public management* (Oxford University Press 2000).
- Hood CC and Margetts HZ, *The tools of government in the digital age* (Palgrave Macmillan 2007).
- Hood C and Peters G, 'The Middle Aging of New Public Management: Into the Age of Paradox?' (2004) 14(3) *J. Public Adm. Res. Theory* 267.
- Hoppe R, *The Governance of Problems: Puzzling, Powering and Participation* (Policy Press 2010).
- Hornung G and Schnabel C, 'Data Protection in Germany I: The Population Census Decision and the Right to Informational Self-Determination' (2009) 25(1) *Computer Law & Security Review* 84 DOI: 10/d2zf3z.
- Huang L, Joseph AD, Nelson B, Rubinstein BIP and Tygar JD, 'Adversarial machine learning' in *Proceedings of the 4th ACM workshop on Security and Artificial Intelligence* (2011) DOI: 10/ft95kn.
- Hullman J, Qiao X, Correll M, Kale A and Kay M, 'In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation' [2018] *IEEE Transactions on Visualization and Computer Graphics* DOI: 10/gfn45q.
- Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES, Spicer K and Wolf P.-P de, *Statistical Disclosure Control* (John Wiley & Sons 2012).

Bibliography

- Hunt M, 'Local Government, Freedom of Expression and Participation' in RA Chapman and M Hunt (eds), *Freedom of information: Local government and accountability* (Ashgate Publishing, Ltd 2010).
- Hutson JA, Taft JG, Barocas S and Levy K, 'Debiasing Desire: Addressing Bias & Discrimination on Intimate Platforms' (2018) 2(CSCW) Proc. ACM Hum.-Comput. Interact. 73:1 DOI: 10/gfkwxv.
- Immergut E, 'Institutions, veto points, and policy results: A comparative analysis of health care' (1990) 10(4) Journal of Public Policy 391 DOI: 10/bvr5nc.
- Information and Privacy Commissioner of Ontario, Canada and Registratiekamer, the Netherlands, *Privacy-Enhancing Technologies: The Path to Anonymity* (Information and Privacy Commissioner and Registratiekamer 1995) (<http://govdocs.ourontario.ca/node/14782>).
- Information Commissioner's Office, *Transparency in outsourcing: a roadmap* (ICO 2015) (<https://ico.org.uk/media/1043531/transparency-in-outsourcing-roadmap.pdf>).
- *Wi-fi location analytics* (ICO 2016) (<https://ico.org.uk/media/for-organisations/documents/1560691/wi-fi-location-analytics-guidance.pdf>).
 - *Big data, artificial intelligence, machine learning and data protection* (ICO 2017) (<https://perma.cc/99ZT-R6TF>).
 - *Feedback Request–Profiling and Automated Decision-Making [v 1.0, 2017/04/06]* (ICO 2017).
 - *In the picture: A data protection code of practice for surveillance cameras and personal information (v1.2)* (ICO 2017) (<https://ico.org.uk/media/1542/cctv-code-of-practice.pdf>).
 - 'The Information Commissioner's Office (ICO) response to DCMS General Data Protection Regulation (GDPR) derogations call for views.' (*ico.org.uk*, 10th May 2017) (<https://perma.cc/33X9-HPHE>).
 - *Democracy Disrupted? Personal Information and Political Influence* (ICO 2018) (<https://perma.cc/2M2N-QQSX>).
 - *Examples of processing 'likely to result in high risk'* (ICO 2018) (<https://perma.cc/SRS6-M7WX>) accessed 28th December 2018.
 - *Investigation into the use of data analytics in political campaigns* (ICO 2018) (<https://perma.cc/2X2U-X6Q4>).
 - *Investigation into the use of data analytics in political campaigns: A report to Parliament* (, ICO 2018).
 - *Technology Strategy, 2018-2021* (ICO 2018) (<https://perma.cc/7RJ5-DAB6>).
- International Working Group on Data Protection in Telecommunications, *Working Paper on Intelligent Video Analytics 58th Meeting, 13-14 October 2015, Berlin (Germany)*-

-
- 675.51.11 (Datenschutz Berlin 2015) (https://www.datenschutz-berlin.de/pdf/publikationen/working-paper/2015/14102015_en_2.pdf).
- Introna L and Wood DM, 'Picturing algorithmic surveillance: The politics of facial recognition systems' (2002) 2(2/3) *Surveillance & Society* DOI: 10/gdxwfx.
- Ipsos MORI, *Public views of Machine Learning* (The Royal Society 2017).
- Jackson S and Forlin G, 'Read My Lips' (2004) 154 *New Law Journal* 1146.
- Janic M, Wijbenga JP and Veugen T, 'Transparency Enhancing Tools (TETs): An Overview' in *Third Workshop on Socio-Technical Aspects in Security and Trust* (2013) DOI: 10/cwmv.
- Japkowicz N and Shah M, *Evaluating learning algorithms: A classification perspective* (Cambridge University Press 2011).
- Ji Z, Lipton ZC and Elkan C, 'Differential Privacy and Machine Learning: a Survey and Review' [2014] arXiv preprint (<https://arxiv.org/abs/1412.7584>).
- Jin Y, Sendhoff B and Körner E, 'Simultaneous Generation of Accurate and Interpretable Neural Network Classifiers' in Y Jin (ed), *Multi-objective machine learning* (2006) DOI: 10/frp7p6.
- Johnson-Williams E, 'TfL needs to give passengers the full picture on WiFi collection scheme' [2016] Open Rights Group Blog (<https://perma.cc/8YEA-BV8D>).
- Johnson JA, 'From Open Data to Information Justice' (2014) 16(4) *Ethics and Information Technology* 263 DOI: 10/gfgt36.
- Jones ML, 'The right to a human in the loop: Political constructions of computer automation and personhood' (2017) 47(2) *Social Studies of Science* 216 DOI: 10/f93mxz.
- Jørgensen TB and Bozeman B, 'Public values: An inventory' (2007) 39(3) *Administration & Society* 354 DOI: 10/ch6kww.
- Jorna F and Wagenaar P, 'The 'Iron Cage' Strengthened? Discretion and Digital Discipline' (2007) 85(1) *Public Admin.* 189 DOI: 10/dqn2m2.
- Kadlec R, Schmid M, Bajgar O and Kleindienst J, 'Text Understanding with the Attention Sum Reader Network' in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics 2016) DOI: 10/cxkt.
- Kahn J, 'To Get Ready for Robot Driving, Some Want to Reprogram Pedestrians' (*Bloomberg*, 16th August 2018) (<https://www.bloomberg.com/news/articles/2018-08-16/to-get-ready-for-robot-driving-some-want-to-reprogram-pedestrians>).
- Kamara I and Kosta E, 'Do Not Track Initiatives: Regaining the Lost User Control' (2016) 6(4) *International Data Privacy Law* 276 DOI: 10/gdxwds.
- Kamiran F and Calders T, 'Data preprocessing techniques for classification without discrimination' (2012) 33(1) *Knowledge and Information Systems* 1 DOI: 10/b36t4t.

- Kamiran F, Calders T and Pechenizkiy M, 'Discrimination aware decision tree learning' in *2010 IEEE International Conference on Data Mining* (2010) DOI: 10/bqdjmp.
- Kasiviswanathan S, Lee H, Nissim K, Raskhodnikova S and Smith A, 'What Can We Learn Privately?' (2011) 40(3) SIAM J. Comput. 793.
- Kasperkevic J, 'Google Says Sorry for Racist Auto-Tag in Photo App' (*The Guardian*, 1st July 2015) (<https://perma.cc/A24K-ZXV6>).
- Kearney RC and Sinha C, 'Professionalism and bureaucratic responsiveness: Conflict or compatibility?' (1988) 48(1) Public Adm. Rev. 571.
- Keats Citron D and Pasquale F, 'The scored society: Due process for automated predictions' (2014) 89(1) Washington Law Review 1.
- 'Building and integrating databases for risk profiles in the United Kingdom', in MS Khwaja, R Awasthi and J Loepnick (eds), *Risk-based tax audits: Approaches and country experiences* (World Bank 2011).
- Kilbertus N, Gascon A, Kusner M, Veale M, Gummadi KP and Weller A, 'Blind Justice: Fairness with Encrypted Sensitive Attributes' in *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)* (2018) (<http://proceedings.mlr.press/v80/kilbertus18a.html>).
- King G, Keohane RO and Verba S, *Designing social inquiry: Scientific inference in qualitative research* (Princeton University Press 1994).
- 'Principle of Equivalence', in W Kirch (ed), *Encyclopedia of Public Health* (Springer 2008) DOI: 10/d698f2.
- 'Principle of Solidarity', in W Kirch (ed), *Encyclopedia of Public Health* (Springer 2008) DOI: 10/bgc69x.
- Kirlappos I, Parkin S and Sasse A, "'Shadow Security" As a Tool for the Learning Organization' (2015) 45(1) SIGCAS Comput. Soc. 29 DOI: 10/gfgrbs.
- Kitchin R, 'Thinking critically about and researching algorithms' (2017) 20(1) Information, Communication & Society 14 DOI: 10/gc3hsj.
- Kleinberg J, Mullainathan S and Raghavan M, 'Inherent Trade-Offs in the Fair Determination of Risk Scores' in CH Papadimitriou (ed), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik 2017) vol 67 DOI: 10/gfgq8s.
- Klimas T and Vaiciukaite J, 'The Law of Recitals in European Community Legislation' (2008) 15(1) ILSA Journal of International and Comparative Law.
- Klitou D, 'Public Space CCTV Microphones and Loudspeakers: The Ears and Mouth of "Big Brother"' in *Privacy-Invasive Technologies and Privacy by Design* (TMC Asser Press 2014) DOI: 10/gfrfdc.

-
- Kohl U, 'Google: the rise and rise of online intermediaries in the governance of the Internet and beyond (Part 2)' (2013) 21(2) *International Journal of Law and Information Technology* 187 DOI: 10.1093/ijlit/eat004.
- Kroll JA, Huey J, Barocas S, Felten EW, Reidenberg JR, Robinson DG and Yu H, 'Accountable Algorithms' (2017) 165 *U Pa. L. Rev.* 633.
- Kuchler H, 'Zuckerberg failed to fix Facebook users' privacy concerns' (*Financial Times*, 17th April 2018) (<https://perma.cc/34EM-DQF7>).
- Kuner C, Cate FH, Millard C and Svantesson DJB, 'The challenge of 'Big Data' for data protection' (2012) 2(2) *International Data Privacy Law* 47 DOI: 10/cvcx.
- Kusner MJ, Loftus J, Russell C and Silva R, 'Counterfactual Fairness' in Guyon, U Luxburg, Bengio, Wallach, Fergus, Vishwanathan and Garnett (eds), *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc 2017) (<http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf>).
- Lacave C and Díez FJ, 'A review of explanation methods for Bayesian networks' (2002) 17(2) *The Knowledge Engineering Review* 107.
- Langheinrich M, 'Privacy by Design – Principles of Privacy-Aware Ubiquitous Systems' in GD Abowd, B Brumitt and S Shafer (eds), *Proceedings of Ubicomp 2001* (Springer 2001) vol 2201.
- Lapsley I, 'New Public Management: The Cruellest Invention of the Human Spirit?' (2009) 45(1) *Abacus* 1.
- Larson J, Mattu S, Kirchner L and Angwin J, 'How We Analyzed the COMPAS Recidivism Algorithm' (*ProPublica*, 23rd May 2016) (<https://perma.cc/W3EB-BKW4>).
- Le Cornu T and Milner B, 'Generating Intelligible Audio Speech from Visual Speech' [2017] (1751) *IEEE/ACM Transactions on Audio, Speech and Language Processing*.
- Leith P, 'Fundamental Errors in Legal Logic Programming' (1986) 29(6) *The Computer Journal* 545 DOI: 10/bzh3hq.
- Lessig L, *Code and Other Laws of Cyberspace* (Basic Books 1999).
- Liberty, 'Liberty's written evidence to the Select Committee on Artificial Intelligence' (*House of Lords Select Committee on Artificial Intelligence*, 1st September 2017) (<https://perma.cc/9LAQ-JWWM>).
- Lim BY and Dey AK, 'Assessing Demand for Intelligibility in Context-Aware Applications' in *Proceedings of the 11th International Conference on Ubiquitous Computing (UbiComp '09, ACM 2009)* DOI: 10/dtkpgv.
- Lim BY, Dey AK and Avrahami D, 'Why and why not explanations improve the intelligibility of context-aware intelligent systems' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)* (ACM 2009) DOI: 10/cq2626.

Bibliography

- Lipsky M, *Street-level bureaucracy: Dilemmas of the individual in public services* (Russell Sage Foundation 2010).
- Lodge M and Wegrich K, *The Problem-solving Capacity of the Modern State: Governance Challenges and Administrative Capacities* (Oxford University Press 2014).
- Lomas N, 'How "anonymous" wifi data can still be a privacy risk' [2017] TechCrunch [⟨https://perma.cc/Y63T-MAC8⟩](https://perma.cc/Y63T-MAC8).
- Loukides G and Shao J, 'Data utility and privacy protection trade-off in k-anonymisation' in *Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society* (ACM 2008).
- Lynskey O, *The Foundations of EU Data Protection Law* (Oxford University Press 2015).
- Lyon D, *Surveillance as social sorting: Privacy, risk, and digital discrimination* (Routledge 2003).
- 'Resisting surveillance', in S Hier and J Greenberg (eds), *The Surveillance Studies Reader* (McGraw-Hill 2007).
- Mahieu RL, Asghari H and van Eeten M, 'Collectively Exercising the Right of Access: Individual Effort, Societal Effect' (2018) 7(3) *Internet Policy Rev.* DOI: 10/cwd8.
- Malgieri G, 'Trade Secrets v Personal Data: a possible solution for balancing rights' (2016) 6(2) *International Data Privacy Law* 102 DOI: 10.1093/idpl/ipv030.
- 'Right to Explanation and Algorithm Legibility in the EU Member States Legislations' [2018] Presented at CPDP 2019, preprint available on SSRN [⟨https://papers.ssrn.com/abstract=3233611⟩](https://papers.ssrn.com/abstract=3233611).
- Malgieri G and Comandé G, 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation' (2017) 7(4) *International Data Privacy Law* 243 DOI: 10/gddkmf.
- Manning CD and Schütze H, *Foundations of Statistical Natural Language Processing* (MIT Press 1999).
- Mantelero A, 'Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection' (2016) 32(2) *Computer Law & Security Review* 238.
- Manzoni J, 'Big data in government: the challenges and opportunities' (*GOVUK*, February 2017) [⟨https://perma.cc/GF7B-5A2R⟩](https://perma.cc/GF7B-5A2R) accessed 4th October 2018.
- Margetts H, *Information Technology in Government: Britain and America* (Routledge 1999).
- Marquess K, 'Redline may be going online' (2000) 86 *ABA J.*
- Marsden CT, *Network Neutrality: From Policy to Law to Regulation* (Manchester University Press 2017) DOI: 10/cxt8.
- 'Prosumer Law and Network Platform Regulation: The Long View towards Creating OffData' (2018) 2(2) *G'town L. Tech. Rev.* 376.

-
- Marsden M, 'ResnetCrowd: A Residual Deep Learning Architecture for Crowd Counting, Violent Behaviour Detection and Crowd Density Level Classification' (*arXiv preprint*, 2017) (<http://arxiv.org/abs/1705.10698>).
- Martens D, Baesens B, Van Gestel T and Vanthienen J, 'Comprehensible credit scoring models using rule extraction from support vector machines' (2007) 183(3) *European Journal of Operational Research* 1466.
- Martin J and others, 'A Study of MAC Address Randomization in Mobile Devices and When it Fails' (2017) 2017(4) *Proceedings on Privacy Enhancing Technologies (PoPETs)* 268 DOI: 10/cwm4.
- Martínez AG, 'Facebook's Not Listening Through Your Phone. It Doesn't Have To' (*Wired*, 11th October 2017) (<https://perma.cc/SLJ6-6YN9>).
- Martinho-Truswell E, 'How AI Could Help the Public Sector' (*Harvard Business Review*, 26th January 2018) (<https://hbr.org/2018/01/how-ai-could-help-the-public-sector>) accessed 1st November 2018.
- Mastelic T, Oleksiak A, Claussen H, Brandic I, Pierson J.-M and Vasilakos AV, 'Cloud Computing: Survey on Energy Efficiency' (2015) 47(2) *ACM CSUR* 33 DOI: 10/cwd6.
- Matus KJ, 'Standardization, certification, and labeling: A background paper for the roundtable on sustainability workshop' in Committee on Certification of Sustainable Products and Services (ed), *Certifiably Sustainable?* (National Academies Press 2010) DOI: 10/gfsjrf.
- Mavroudis V and Veale M, 'Eavesdropping Whilst You're Shopping: Balancing Personalisation and Privacy in Connected Retail Spaces' in *Proceedings of the 2018 PET-RAS/IoTUK/IET Living in the IoT Conference* (IET 2018) DOI: 10/gffng2.
- Mayer-Schönberger V and Cukier K, *Big Data: A Revolution That Will Transform How We Live, Work and Think* (Hodder & Stoughton 2013).
- Mazzucato M, *The entrepreneurial state: Debunking public vs. private sector myths* (Anthem Press 2015).
- McDuff D, El Kaliouby R, Kodra E and Picard R, 'Measuring Voter's Candidate Preference Based on Affective Responses to Election Debates' in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (2013) DOI: 10/gfrfc9.
- McKeown B and Thomas DB, *Q Methodology* (SAGE 2013).
- McNamee J, 'Is Privacy Still Relevant in a World of Bastard Data?' (*EDRI editorial*) (<https://edri.org/enditorial-is-privacy-still-relevant-in-a-world-of-bastard-data>).
- McSherry F, 'Statistical inference considered harmful' [2016] (<https://github.com/frankmsherry/blog/blob/master/posts/2016-06-14.md>).
- Meadows D, 'Dancing with systems' (2002) 13(2) *The Systems Thinker* (<https://perma.cc/4RPZ-XN3H>).

Bibliography

- Meadows DH, *Thinking in systems* (Earthscan 2008).
- 'measure', in *Oxford English Dictionary* (2nd edn, Oxford University Press 2014) (<http://www.oed.com/view/Entry/115506>).
- Mendoza I and Bygrave LA, 'The Right Not to Be Subject to Automated Decisions Based on Profiling' in T.-E Synodinou, P Jougleux, C Markou and T Prastitou (eds), *EU Internet Law* (Springer 2017) DOI: 10/gfscwg.
- Menon N, 'Universalism without Foundations?' (2002) 31(1) *Economy and Society* 152 DOI: 10/fsc89q.
- Merken S, 'Irish Twitter Probe seen as Test Case for EU Privacy Rules' (*Bloomberg Law*, 18th October 2018) (<https://perma.cc/M83P-HX27>).
- Merry SE, *The seductions of quantification: Measuring human rights, gender violence, and sex trafficking* (University of Chicago Press 2016).
- Meystre SM, Friedlin FJ, South BR, Shen S and Samore MH, 'Automatic de-identification of textual documents in the electronic health record: a review of recent research' (2010) 10(1) *BMC Medical Research Methodology* 70.
- Microsoft, 'Azure Cognitive Services Emotion API Documentation' (*Microsoft Azure*, 27th June 2017) (<https://perma.cc/BC6Z-FY78>).
- Mikolov T, Chen K, Corrado G and Dean J, 'Efficient Estimation of Word Representations in Vector Space' [2013] arXiv preprint (<https://arxiv.org/abs/1301.3781>).
- Mikolov T, Sutskever I, Chen K, Corrado GS and Dean J, 'Distributed Representations of Words and Phrases and their Compositionality' in CC Burges, Bottou, Welling, Ghahramani and K Weinberger (eds), *Advances in Neural Information Processing Systems 26* (2013).
- Mikolov T, Yih W.-t and Zweig G, 'Linguistic Regularities in Continuous Space Word Representations' in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2013).
- Miller P and O'Leary T, 'Accounting and the construction of the governable person' (1987) 12(3) *Accounting, Organizations and Society* 235 DOI: 10/d3cg8w.
- Mills A, 'An Update on Consultation' (2015) 20(3) *Judicial Review* 160 DOI: 10/gfsjqs.
- Milner C and Berg B, *Tax Analytics: Artificial Intelligence and Machine Learning* (PwC Advanced Tax Analytics & Innovation 2017) (<https://perma.cc/4TW3-5P8N>).
- Minsky M and Papert S, *Perceptrons* (MIT Press 1969).
- Mironov I, 'On Significance of the Least Significant Bits for Differential Privacy' in *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS'12)* (New York, NY, USA, 2012) DOI: 10/cwjw.

-
- Missier P and others, 'Linking multiple workflow provenance traces for interoperable collaborative science' in *The 5th Workshop on Workflows in Support of Large-Scale Science* (IEEE 2010) DOI: 10/dsxdf6.
- Mitchell TM, *Machine learning* (McGraw Hill 1997).
- Mittelstadt B, 'From Individual to Group Privacy in Big Data Analytics' (2017) 30(4) *Philos. Technol.* 475 DOI: 10/cwdg.
- Mittelstadt BD, Allo P, Taddeo M, Wachter S and Floridi L, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3(2) *Big Data & Society* 2053951716679679 DOI: 10/gcdx92.
- Mizroch A, 'Artificial-intelligence experts are in high demand' (*The Wall Street Journal*, 1st May 2015).
- Montavon G, Lapuschkin S, Binder A, Samek W and Müller K.-R, 'Explaining nonlinear classification decisions with deep Taylor decomposition' (2017) 65 *Pattern Recognition* 211 DOI: 10/f9vv35.
- Montjoye Y.-A de, Hidalgo CA, Verleysen M and Blondel VD, 'Unique in the Crowd: The privacy bounds of human mobility' (2013) 3 *Sci. Rep.* 1376 DOI: 10/msd.
- Moore JD and Swartout WR, 'A reactive approach to explanation: Taking the user's feedback into account' in C Paris, WR Swartout and WC Mann (eds), *Natural Language Generation in Artificial Intelligence and Computational Linguistics* (Springer 1991).
- Moore MH, *Creating public value: Strategic management in government* (Harvard University Press 1995).
- R Moore (ed), *A compendium of research and analysis on the Offender Assessment System* (Ministry of Justice Analytical Series 2015) (<https://perma.cc/W2FT-NFWZ>).
- 'More accountability for big-data algorithms' (2016) 537(7621) *Nature* 449 DOI: 10/gfgrbj.
- Morgenthaler J, Gridnev M, Sauciuc R and Bhansali S, 'Searching for Build Debt: Experiences Managing Technical Debt at Google' in *Proceedings of the Third International Workshop on Managing Technical Debt (MTD '12)* (IEEE Press 2012).
- Munk TB, '100,000 False Positives for Every Real Terrorist: Why Anti-Terror Algorithms Don't Work' (2017) 22(9) *First Monday* DOI: 10/cvzf.
- Munroe R, 'Self Driving' (*xkcd webcomic*, 2017) (<https://xkcd.com/1897/>).
- Muralidhar K and Sarathy R, 'Does Differential Privacy Protect Terry Gross' Privacy?' in J Domingo-Ferrer and E Magkos (eds), *Proceedings of Privacy in Statistical Databases (PSD 2010)* (Springer 2010) DOI: 10/cr2zq7.
- National Science and Technology Council, *Preparing for the future of artificial intelligence* (US Government 2016) (<https://perma.cc/DDR3-2QBH>).

Bibliography

- Neches R, Swartout WR and Moore JD, 'Enhanced maintenance and explanation of expert systems through explicit models of their development' [1985] (11) IEEE Transactions on Software Engineering 1337.
- Neff G and Nafus D, *Self-Tracking* (MIT Press 2016).
- Neuman GL, 'Subsidiarity' in *The Oxford Handbook of International Human Rights Law* (Oxford University Press 2013) DOI: 10/gfrjh9.
- Nguyen FD, 'Regulation of Medical Expert Systems: A Necessary Evil' (1993) 34 Santa Clara L. Rev. 1187.
- Ni Loideain N, 'A Bridge too Far? The Investigatory Powers Act 2016 and Human Rights Law' in L Edwards (ed), *Law, Policy and the Internet* (Hart 2018).
- Nissim K and others, *Differential Privacy: A Primer for a Non-technical Audience* (A product of the "Bridging Privacy Definitions" working group, part of the Privacy Tools for Sharing Research Data project at Harvard University 2017).
- Noble SU, *Algorithms of Oppression* (NYU Press 2018).
- Northpointe, Incd/b/a equivant, 'Official Response to Science Advances' (*Equivant (press release)*, 17th January 2018) (<http://perma.cc/YB6F-PZW9>).
- Northrop A, Kraemer KL, Dunkle D and King JL, 'Payoffs from Computerization: Lessons over Time' (1990) 50(5) Public Adm. Rev. 505 DOI: 10.2307/976781.
- Norton PD, *Fighting Traffic: The Dawn of the Motor Age in the American City* (MIT Press 2008).
- Nussbaum MC, *Women and human development: The capabilities approach* (Cambridge University Press 2001).
- O'Hara K, 'The Seven Veils of Privacy' (2016) 20(2) IEEE Internet Computing 86 DOI: 10/gfbzpw.
- O'Neil C, *Weapons of Math Destruction* (Penguin 2016).
- Odendahl T and Shaw AM, 'Interviewing elites' in J Gubrium and J Holstein (eds), *Handbook of Interview Research* (SAGE 2002) DOI: 10/gfgq8h.
- Ohm P, 'Broken promises of privacy: Responding to the surprising failure of anonymization' (2009) 57 UCLA L. Rev. 1701.
- Okin SM, 'Poverty, Well-Being, and Gender: What Counts, Who's Heard?' (2005) 31(3) Philosophy & Public Affairs 280 DOI: 10/cr3hqq.
- Olejnik L, 'Privacy of London Tube Wifi Tracking' (*Security, Privacy & Tech Inquiries [blog]*) (<https://blog.lukaszolejnik.com/privacy-of-london-tube-wifi-tracking/>).
- Olson M, *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard University Press 1965).
- OpinionLab, New study: consumers overwhelmingly reject in-store tracking by retailers (March 2014) (<https://perma.cc/R289-HZ3Z>).

-
- Oppenheimer M, 'Reclaiming 'Jew'' (*The New York Times*, 22nd April 2017) (<https://perma.cc/8QM2-DHDB>).
- Orlowski A, 'Apple pollutes data about you to protect your privacy. But it might not be enough' [2016] *The Register* (<https://perma.cc/98SX-CTTR>).
- Oswald M, 'Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power' (2018) 376(2128) *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* DOI: 10/gdxt27.
- Oswald M and Grace J, 'Intelligence, policing and the use of algorithmic analysis: A freedom of information-based study' (2016) 1(1) *Journal of Information Rights, Policy and Practice* DOI: 10.21039/irpandp.v1i1.16.
- Oswald M, Grace J, Urwin S and Barnes GC, 'Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and 'Experimental' Proportionality' (2018) 27(2) *Information & Communications Technology Law* 223 DOI: 10/gdz288.
- Overdorf R, Kulynych B, Balsa E, Troncoso C and Gürses S, 'POTs: Protective Optimization Technologies' [2018] arXiv preprint (<https://arxiv.org/abs/1806.02711>).
- Oye KA and others, 'Regulating Gene Drives' (2014) 345(6197) *Science* 626 DOI: 10/gfphx3.
- Padmanabhan J and Johnson Premkumar MJ, 'Machine learning in automatic speech recognition: A survey' (2015) 32(4) *IETE Technical Review* 240.
- Page EC and Jenkins B, *Policy bureaucracy: Government with a cast of thousands* (Oxford University Press 2005).
- Pasquale F, 'Beyond innovation and competition: The need for qualified transparency in internet intermediaries' (2010) 104(1) *Nw. U. L. Rev.*
- *The Black Box Society: The Secret Algorithms that Control Money and Information* (Harvard University Press 2015).
- Pearl J and Mackenzie D, *The Book of Why: The New Science of Cause and Effect* (Basic Books 2018).
- Pedreschi D, Ruggieri S and Turini F, 'Discrimination-aware data mining' in *ACM KDD '08* (ACM 2008) DOI: 10/c7xx96.
- Pellissier Tanon T, Vrandečić D, Schaffert S, Steiner T and Pintscher L, 'From Freebase to Wikidata: The Great Migration' in *Proceedings of the 25th International Conference on World Wide Web (WWW 2016)* (International World Wide Web Conferences Steering Committee 2016).
- Pennington J, Socher R and Manning C, 'GloVe: Global Vectors for Word Representation' in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014).

Bibliography

- Pennington J, Socher R and Manning C, 'GloVe: Global vectors for word representation' in *Empirical Methods in Natural Language Processing (EMNLP)* (2014) (<http://www.aclweb.org/anthology/D14-1162>).
- Perrow C, *Normal accidents: Living with high risk technologies* (Basic Books 1984).
- Perry WL, McInnis B, Price CC, Smith SC and Hollywood JS, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (RAND Corporation 2013) (http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf).
- Petajan ED, 'Automatic Lipreading to Enhance Speech Recognition' (Doctoral dissertation, University of Illinois at Urbana-Champaign 1984).
- Pfitzmann A and Köhntopp M, 'Anonymity, Unobservability, and Pseudonymity – A Proposal for Terminology' in *Designing Privacy Enhancing Technologies* (Springer 2001).
- Phelps ES, 'The Statistical Theory of Racism and Sexism' (1972) 62(4) *The American Economic Review* 659 (<https://www.jstor.org/stable/1806107>).
- Phillips N, 'Opening Keynote of Commissioner Noah Joshua Phillips, Washington, DC' (*US Federal Trade Commission*, 6th February 2019) (https://www.ftc.gov/system/files/documents/public_statements/1452828/phillips_-_fpf_opening_keynote_2-6-19.pdf).
- Picard RW, *Affective Computing* (The MIT Press 1997).
- Pigou A, *The Economics of Welfare* (first published 1952, Routledge 2017) DOI: 10/gfgq79.
- Pleiss G, Raghavan M, Wu F, Kleinberg J and Weinberger KQ, 'On Fairness and Calibration' in Guyon, U Luxburg, Bengio, Wallach, Fergus, Vishwanathan and Garnett (eds), *Advances in Neural Information Processing Systems 30* (Curran Associates, Inc 2017).
- Pollock S, 'Twitter faces investigation by privacy watchdog over user tracking' *The Sunday Times* (London, 21st October 2018).
- Poon M, 'From new deal institutions to capital markets: Commercial consumer risk scores and the making of subprime mortgage finance' (2009) 34(5) *Accounting, Organizations and Society* 654 DOI: 10/cm8g3x.
- Posa JG, 'Smart Phone with Self-Training, Lip-Reading and Eye-Tracking Capabilities' (*US Patent no 20130332160*) (<https://patents.google.com/patent/US20130332160>).
- Potamianos G, Neti C, Gravier G and Garg A, 'Recent Advances in the Automatic Recognition of Audio-Visual Speech' (2003) 91(1306) *Proceedings of the IEEE* DOI: 10.1109/JPROC.2003.817150.
- Prakken H, *Logical tools for modelling legal argument* (Kluwer 1997).

-
- Purtova N, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10(1) Law, Innovation and Technology 40 DOI: 10/gd4rmh.
- Pyrgelis A, Troncoso C and De Cristofaro E, 'Knock Knock, Who's There? Membership Inference on Aggregate Location Data' in *Proceedings of the 25th Network and Distributed System Security Symposium (NDSS 2018)* (2018) (<https://arxiv.org/abs/1708.06145>).
- E Quill and RJ Friel (eds), *Damages and Compensation Culture: Comparative Perspectives* (Bloomsbury Publishing 2016).
- Quiñonero-Candela J, Sugiyama M, Schwaighofer A and Lawrence ND, *Dataset shift in machine learning* (The MIT Press 2009).
- Quisquater J.-J, Guillou L, Annick M and Berson T, 'How to Explain Zero-knowledge Protocols to Your Children' in *Proceedings on Advances in Cryptology (CRYPTO '89)* (Springer 1989).
- Radford A, Wu J, Child R, Luan D, Amodei D and Sutskever I, 'Language Models Are Unsupervised Multitask Learners' [2019] OpenAI Working Paper.
- Rauhofer J, 'Of Men and Mice: Should the EU Data Protection Authorities' Reaction to Google's New Privacy Policy Raise Concern for the Future of the Purpose Limitation Principle' (2015) 1 European Data Protection Law Review (EDPL) 5 DOI: 10/gfsh4s.
- Rayner K, Carlson M and Frazier L, 'The Interaction of Syntax and Semantics during Sentence Processing: Eye Movements in the Analysis of Semantically Biased Sentences' (1983) 22(358) Journal of Verbal Learning and Verbal Behaviour.
- Reisman D, Schultz J, Crawford K and Whittaker M, *Algorithmic Impact Assessments* (AI Now Institute 2018) (<https://perma.cc/H79W-JN8F>).
- Rekdal OB, 'Academic Urban Legends' (2014) 44(4) Social Studies of Science 638 DOI: 10/gd89bc.
- Resnick B, 'How Artificial Intelligence Learns to Be Racist' [2017] Vox (<https://perma.cc/4ZJP-HB3G>) accessed 2nd October 2018.
- Ribeiro MT, Singh S and Guestrin C, "Why should I trust you?": Explaining the predictions of any classifier' in (2016) DOI: 10/gfgrbd.
- 'Model-Agnostic Interpretability of Machine Learning', in *Presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY* (2016) (<https://arxiv.org/abs/1606.05386>).
- Robinson DG and Bogen M, *Automation & the Quantified Society* (Upturn and the Netgain Partnership 2017) (<https://perma.cc/QB98-ARFF>).
- Rock F, 'Policy and Practice in the Anonymisation of Linguistic Data' (2001) 6(1) International Journal of Corpus Linguistics 1 DOI: 10/cg534s.

Bibliography

- Rosenblat A, Levy KE, Barocas S and Hwang T, 'Discriminating Tastes: Uber's Customer Ratings as Vehicles for Workplace Discrimination' (2017) 9(3) Policy & Internet 256 DOI: 10/gddxqn.
- Rottenburg R, Merry SE, Park S.-J and Mugler J, *The world of indicators: The making of governmental knowledge through quantification* (Cambridge University Press 2015).
- Rouvroy A, 'Technology, virtuality and utopia' in M Hildebrandt and A Rouvroy (eds), *Law, Human Agency and Autonomic Computing* (Routledge 2011).
- Royal United Services Institute for Defence and Security Studies (RUSI), *Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges* (, RUSI 2018).
- Rubinstein IS, 'Big Data: The End of Privacy or a New Beginning?' (2013) 3(2) International Data Privacy Law 74 DOI: 10/cvcz.
- Rudin C, 'Please Stop Explaining Black Box Models for High Stakes Decisions' in *NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning* (2018) (<https://arxiv.org/abs/1811.10154>).
- Ruiz J and Johnson-Williams E, *Debates, awareness, and projects about GDPR and data protection: Interim Report for the Information Commissioner's Office for the project: "Making new privacy rights protect and enable people's financial futures"* (Open Rights Group and Projects by If 2018).
- Sadeghi A.-R, Schneider T and Wehrenberg I, 'Efficient Privacy-Preserving Face Recognition' in D Lee and S Hong (eds), *Information, Security and Cryptology – ICISC 2009* (Lecture Notes in Computer Science, Springer Berlin Heidelberg 2010).
- Salton G, Wong A and Yang C, 'A Vector Space Model for Automatic Indexing' (1975) 18(11) Commun. ACM 613 DOI: 10/fw8vv8.
- Sánchez D and Batet M, 'C-Sanitized: A Privacy Model for Document Redaction and Sanitization' (2016) 67(1) Journal of the Association for Information Science and Technology 148 DOI: 10/f77swg.
- 'Toward Sensitive Document Release with Privacy Guarantees' (2017) 59 Engineering Applications of Artificial Intelligence 23 DOI: 10/f9s634.
 - 'Toward sensitive document release with privacy guarantees' (2017) 59 Engineering Applications of Artificial Intelligence 23.
- Sánchez D, Batet M and Viejo A, 'Minimizing the Disclosure Risk of Semantic Correlations in Document Sanitization' (2013) 249 Information Sciences 110 DOI: 10/f5b2pn.
- Sandvig C, Hamilton K, Karahalios K and Langbort C, 'Auditing algorithms: Research methods for detecting discrimination on internet platforms' in *Data and Discrimin-*

-
- ation: *Converting Critical Concerns into Productive Inquiry* (Seattle, WA, 2014) (<https://perma.cc/8JKK-FMUV>).
- Saunders B, Kitzinger J and Kitzinger C, 'Anonymising Interview Data: Challenges and Compromise in Practice' (2015) 15(5) *Qualitative Research* 616 DOI: 10/f7sk3r.
- Schreurs W, Hildebrandt M, Kindt E and Vanfleteren M, 'Cogitas, Ergo Sum. The Role of Data Protection Law and Non-discrimination Law in Group Profiling in the Private Sector' in M Hildebrandt and S Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer 2008) DOI: 10/ccpqh5.
- Science and Technology Committee (Commons), *Algorithms in Decision-Making* (HC 2018, 351) (<https://perma.cc/PH52-NUWC>).
- Scott C, 'Enforcing Consumer Protection Laws' in G Howells, I Ramsay and T Wilhelmsen (eds), *Handbook of Research on International Consumer Law, Second Edition* (Edward Elgar 2018) DOI: 10.4337/9781785368219.
- Scott JC, *Seeing like a state: How certain schemes to improve the human condition have failed* (Yale University Press 1998).
- Scott M, 'How Big Tech Learned to Love Regulation' (*POLITICO*, 11th November 2018) (<https://perma.cc/CQ5N-J3Q9>).
- Sculley D and others, 'Hidden technical debt in machine learning systems' in *Proceedings of the 28th International Conference on Neural Information Processing Systems, Montréal, Canada – December 07 - 12, 2015* (Cambridge, MA, 2015).
- Seaver N, 'Knowing Algorithms' in *Paper presented at Media in Transition 8, Cambridge, MA* (2013).
- 'Algorithms as Culture: Some tactics for the ethnography of algorithmic systems' (2017) 4(2) *Big Data & Society* DOI: 10/gd8fdx.
- Sebe ADJKSRGN, 'The More the Merrier: Analysing the Affect of a Group of People in Images' in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (IEEE 2015) DOI: 10/gfrfdf.
- Secretary-General of the United Nations, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/73/150)* (United Nations 2018).
- Selbst AD and Powles J, 'Meaningful information and the right to explanation' (2017) 7(4) *International Data Privacy Law* 233 DOI: 10/gddxmz.
- Selbst A and Barocas S, 'The Intuitive Appeal of Explainable Machines' (2018) 87 *Fordham L. Rev.* 1085 DOI: 10/gdz285.
- Select Committee on Artificial Intelligence (Lords), *AI in the UK: ready, willing and able?* (HL 2018, Paper).
- Sen AK, *Development as freedom* (Oxford University Press 1999).

- Shadbolt NR, Smith DA, Simperl E, Van Kleek M, Yang Y and Hall W, 'Towards a classification framework for social machines' in *Proceedings of the 22nd International Conference on World Wide Web* (2013) DOI: 10/gfgq9b.
- Shao J, Kang K, Loy CC and Wang X, 'Deeply Learned Attributes for Crowded Scene Understanding' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)* (IEEE 2015) DOI: 10/gfrfc7.
- Sharif M, Bhagavatula S, Bauer L and Reiter MK, 'Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition' in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016) DOI: 10/bsmm.
- Sharma A and Kumar Panigrahi P, 'A review of financial accounting fraud detection based on data mining techniques' (2012) 39(1) IJCAI 37.
- Shneiderman B and Maes P, 'Direct manipulation vs. interface agents' (1997) 4(6) Interactions 42.
- Shokri R, Stronati M, Song C and Shmatikov V, 'Membership Inference Attacks Against Machine Learning Models' in *2017 IEEE Symposium on Security and Privacy (SP)* (2017) DOI: 10/cwdq.
- Shortliffe E, *Computer-Based Medical Consultations: MYCIN* (Elsevier 1976).
- Simonite T, 'When It Comes to Gorillas, Google Photos Remains Blind' (*WIRED*, 11th January 2018) (<https://perma.cc/L52Z-J8J6>).
- Simperl E and Luczak-Rösch M, 'Collaborative ontology engineering: A survey' (2014) 29(1) Knowl. Engin. Rev. 101.
- Singh J, Cobbe J and Norval C, 'Decision Provenance: Capturing data flow for accountable systems' [2018] arXiv preprint (<https://arxiv.org/abs/1804.05741>).
- Skitka LJ, Mosier KL and Burdick M, 'Does automation bias decision-making?' (1999) 51 International Journal of Human-Computer Studies 991 DOI: 10/bg5rb7.
- Sloot B van der, 'From Data Minimization to Data Minimummization' in B Custers, T Calders, B Schermer and T Zarsky (eds), *Discrimination and privacy in the information society* (Springer 2012) DOI: 10/cwqq.
- Smith ML, Noorman ME and Martin AK, 'Automating the public sector and organizing accountabilities' (2010) 26(1) Communications of the Association for Information Systems.
- Solon O, 'The Rise of 'Pseudo-AI': How Tech Firms Quietly Use Humans to Do Bots' Work' (*The Guardian*, 6th July 2018) (<https://perma.cc/Q4P9-ZCHF>).
- Solove DJ, 'Introduction: Privacy Self-Management and the Consent Dilemma' (2012) 126 Harvard Law Review 1880.

-
- Song C, Ristenpart T and Shmatikov V, 'Machine Learning Models that Remember Too Much' in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)* (2017) DOI: 10/cwdp.
- Spradling C, Soh L.-K and Ansoorge C, 'Ethics training and decision-making: do computer science programs need help?' in *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education* (ACM 2008) vol 40.
- Sridhar V, Subramanian S, Arteaga D, Sundararaman S, Roselli D and Talagala N, 'Model Governance: Reducing the Anarchy of Production ML' in *USENIX ATC'18* (USENIX Association 2018).
- Stanton JM, 'Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors' (2001) 9(3) *Journal of Statistics Education* DOI: 10/gd82dx.
- Star SL and Griesemer JR, 'Institutional Ecology, Translations and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39' (1989) 19(3) *Social Studies of Science* 387 DOI: 10/ckpxb6.
- State Council of the People's Republic of China, *A Next Generation Artificial Intelligence Development Plan* (Creemers R, Webster G, Tsai P and Kania E trs, Government of China 2017) (<https://perma.cc/9EE3-4MXH>).
- Statistics New Zealand, 'Integrated Data Infrastructure' (*Government of New Zealand*, 2016) (<https://perma.cc/9RXL-SV7P>) accessed 4th October 2018.
- Stepanek M, 'Weblining: Companies are using your personal data to limit your choices—and force you to pay more for products' [2000] *Bloomberg Business Week*.
- Stephan K, 'Apple Versus the Feds: How a smartphone stymied the FBI' (2017) 6(2) *IEEE Consumer Electronics Mag.* 103 DOI: 10/cwfp.
- Sterman JD, *Business Dynamics* (McGraw-Hill 2000).
- Stilgoe J, *Experiment Earth: Responsible Innovation in Geoengineering* (Earthscan 2015).
- Stilgoe J, Owen R and Macnaghten P, 'Developing a framework for responsible innovation' (2013) 42(9) *Research Policy* 1568 DOI: 10/f5gv8h.
- Stivers C, 'The Listening Bureaucrat: Responsiveness in Public Administration' (1994) 54(4) *Public Adm. Rev.* 364 DOI: 10/dr39pq.
- Strong SI, 'From Class to Collective: The De-Americanization of Class Arbitration' (2010) 26(4) *Arbitration International* 493 DOI: 10/gfr9hf.
- Stucke ME and Grunes AP, 'No Mistake About It: The Important Role of Antitrust in the Era of Big Data' [2015] *U Tennessee Legal Studies Research Paper*.
- Sullivan D, 'Google In Controversy Over Top-Ranking For Anti-Jewish Site' (*Search Engine Watch*, 24th April 2004) (<https://perma.cc/8ZCC-7WFB>).
- Susskind R, *Expert Systems in Law* (Clarendon Press 1987).

- Sutcu Y, Sencar HT and Memon N, 'A Secure Biometric Authentication Scheme Based on Robust Hashing' in *Proceedings of the 7th Workshop on Multimedia and Security* (ACM 2005) DOI: 10/cj8sj5.
- Suwajanakorn S, Seitz SM and Kemelmacher-Shlizerman I, 'Synthesizing Obama: Learning Lip Sync from Audio' (2017) 36(4) ACM Trans. Graph. 95:1 DOI: 10/gdgp4.
- Swallow D and Bourke G, *The Freedom of Information Act and Higher Education: The experience of FOI officers in the UK* (The Constitution Unit, University College London June 2012) (<https://www.ucl.ac.uk/constitution-unit/research/foi/foi-universities/he-foi-officers-survey-report.pdf>).
- Swartout WR, 'XPLAIN: A system for creating and explaining expert consulting programs' (1983) 21(3) Artificial Intelligence 285 DOI: 10/bxgmhx.
- Sweeney L, 'Simple Demographics Often Identify People Uniquely' [2000] Carnegie Mellon Univ. Data Privacy Working Paper 3 (<https://perma.cc/6V9N-9QNG>).
- 'Discrimination in Online Ad Delivery' (2013) 11(3) Queue 10 DOI: 10/gdxwj6.
 - 'Saving Humanity', in *1st Conference on Fairness, Accountability and Transparency (FAT*)*, New York, 23 February (Keynote address) (2018) (https://youtu.be/OlK_nVOM2tc).
- Swindon P, 'Lip-Reading CCTV Set to Capture Shoppers' Private Comments for Big Companies' (*Sunday Herald (Scotland)*, 17th August 2017) (<http://perma.cc/U5KH-XG9A>).
- Tan J, Wang X, Nguyen C.-T and Shi Y, 'SilentKey: A New Authentication Framework Through Ultrasonic-Based Lip Reading' (2018) 2(1) Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 36:1 DOI: 10/gdz6qh.
- Tang J, Korolova A, Bai X, Wang X and Wang X, 'Privacy Loss in Apple's Implementation of Differential Privacy on macOS 10.12' [2017] arXiv preprint (<https://arxiv.org/abs/1709.02753>).
- Taylor C, *Multiculturalism: Examining the Politics of Recognition* (Princeton University Press 1994).
- Taylor L, 'What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally' (2017) 4(2) Big Data & Society 2053951717736335 DOI: 10/gfgt4b.
- Taylor L, Floridi L and Sloot B van der, *Group Privacy* (Springer 2017).
- Teach RL and Shortliffe EH, 'An analysis of physician attitudes regarding computer-based clinical consultation systems' (1981) 14(6) Computers and Biomedical Research 542.
- The Law Society of England and Wales, *Legal aid deserts in England and Wales* (2016) (<https://perma.cc/268A-5TJM>).
- The Royal Society, *Machine learning: The power and promise of computers that learn by example* (The Royal Society 2017).

-
- The Royal Society and the British Academy, *Data management and use: Governance in the 21st Century* (The Royal Society and the British Academy 2017).
- Theodorou L, Healey PG and Smeraldi F, 'Exploring Audience Behaviour During Contemporary Dance Performances' in *Proceedings of the 3rd International Symposium on Movement and Computing* (ACM 2016).
- Thiemann A, Gonzaga P and Stucke ME, *DAF/COMP(2016)14: Big Data: Bringing competition policy to the digital era - Background note by the Secretariat* (Paris, 201) ([https://one.oecd.org/document/DAF/COMP\(2016\)14/en/pdf](https://one.oecd.org/document/DAF/COMP(2016)14/en/pdf)).
- Thies J, Zollhofer M, Stamminger M, Theobalt C and Niessner M, 'Face2Face: Real-Time Face Capture and Reenactment of RGB Videos' in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
- Thomson ME, Önkal D, Avcioglu A and Goodwin P, 'Aviation risk perception: A comparison between experts and novices' (2004) 24(6) *Risk Analysis* 1585 DOI: 10.1111/j.0272-4332.2004.00552.x.
- Tibshirani R, 'Regression Shrinkage and Selection via the Lasso' (1996) 58(1) *Journal of the Royal Statistical Society. Series B (Methodological)* 267.
- Tickle AB, Andrews R, Golea M and Diederich J, 'The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks' (1998) 9(6) *IEEE Transactions on Neural Networks* 1057 DOI: 10/btn5vv.
- Tintarev N and Masthoff J, 'Designing and evaluating explanations for recommender systems' in *Recommender Systems Handbook* (Springer 2011).
- Tjong Tjin Tai E, 'Liability for (Semi)Autonomous Systems: Robots and Algorithms' in V Mak, E Tjong Tjin Tai and A Berlee (eds), *Research Handbook on Data Science and Law* (Edward Elgar 2018) DOI: 10/gfsq6x.
- Tollenaar N, Wartna BJ, Van Der Heijden PM and Bogaerts S, 'StatRec – Performance, validation and preservability of a static risk prediction instrument' (2016) 129(1) *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 25 DOI: 10/gfgrbb.
- Tramèr F, Zhang F, Juels A, Reiter MK and Ristenpart T, 'Stealing Machine Learning Models via Prediction APIs' in *USENIX Security Symposium* (2016).
- Transport for London, *Insights from Wi-Fi Data: Proposed Pilot* (TfL (Released under the Freedom of Information Act 2000) 2016) (<https://perma.cc/NSS3-7RW5>).
- *TfL WiFi Analytics Briefing Pack* (November 2016) (<https://perma.cc/7PHN-WBGH>).
- *Review of the TfL WiFi Pilot* (TfL 2017).

- Tsebelis G, 'Decision making in political systems: Veto players in presidentialism, parliamentarism, multicameralism and multipartyism' (1995) 25(3) *British Journal of Political Science* 289 DOI: 10/cc622q.
- Tukey JW, 'We need both exploratory and confirmatory' (1980) 34(1) *Am. Stat.* 23.
- Turney P and Pantel P, 'From Frequency to Meaning: Vector Space Models of Semantics' (2010) 37 *Journal of Artificial Intelligence Research* 141 DOI: 10/gd85zk.
- Tutt A, 'An FDA for Algorithms' (2017) 69 *Administrative Law Review* 83.
- Tversky A and Kahneman D, 'Judgment under Uncertainty: Heuristics and Biases' (1974) 185(4157) *Science* 1124.
- Tzanou M, 'Data Protection as a Fundamental Right next to Privacy? 'Reconstructing' a Not so New Right' (2013) 3(2) *International Data Privacy Law* 88 DOI: 10/gfbzpf.
- United Nations Special Rapporteur on Extreme Poverty, *Statement on Visit to the United Kingdom, by Professor Philip Alston, United Nations Special Rapporteur on extreme poverty and human rights* (2018) (https://www.ohchr.org/Documents/Issues/Poverty/EOM_GB_16Nov2018.pdf).
- Ursic H, 'Unfolding the New-Born Right to Data Portability: Four Gateways to Data Subject Control' (2018) 15(1) *SCRIPTed* 42 DOI: 10/gfc7c9.
- Ustun B and Rudin C, 'Supersparse Linear Integer Models for Optimized Medical Scoring Systems' (2016) 102(3) *Machine Learning* 349 DOI: 10/f8crhw.
- Ustun B, Spangher A and Liu Y, 'Actionable Recourse in Linear Classification' in *Proceedings of the ACM Conference on Fairness, Accountability and Transparency (ACM FAT* 2019)* (ACM 2019) (<http://arxiv.org/abs/1809.06514>).
- Vaas L (*Sophos Naked Security*) (<https://nakedsecurity.sophos.com/2017/05/11/would-you-like-a-side-of-facial-recognition-with-your-pizza/>).
- Van der Wal Z, De Graaf G and Lasthuizen K, 'What's valued most? Similarities and differences between the organizational values of the public and private sector' (2008) 86(2) *Public Administration* 465 DOI: 10/bwjt35.
- Van Kleek M, Binns R, Zhao J, Slack A, Lee S, Ottewell D and Shadbolt N, 'X-Ray Refine: Supporting the exploration and refinement of information exposure resulting from smartphone apps' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'18)* (ACM 2018) DOI: 10/evcn.
- Van Kleek M, Liccardi I, Binns R, Zhao J, Weitzner DJ and Shadbolt N, 'Better the Devil You Know: Exposing the Data Sharing Practices of Smartphone Apps' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'17)* (ACM 2017) DOI: 10/gfgq8q.
- Van Kleek M, Seymour W, Veale M, Binns R and Shadbolt N, 'The Need for Sensemaking in Networked Privacy and Algorithmic Responsibility' in *Sensemaking in a*

-
- Senseless World: Workshop at ACM CHI'18, 22 April 2018, Montréal, Canada* (2018) (<http://discovery.ucl.ac.uk/id/eprint/10046886>).
- Vanhoef M, Matte C, Cunche M, Cardoso LS and Piessens F, 'Why MAC Address Randomization is Not Enough: An Analysis of Wi-Fi Network Discovery Mechanisms' in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security (ASIACCS'16)* (ACM 2016).
- Various, 'OpenReview Page for ICLR 2017 Submitted Article: "LipNet: End-to-End Sentence-Level Lipreading"' (*OpenReview*, 1st February 2017) (<https://openreview.net/forum?id=BkjLkSxqg¬eId=BkjLkSxqg>).
- Veale M, *Data management and use: case studies of technologies and governance* (The Royal Society and the British Academy 2017).
- 'Logics and Practices of Transparency and Opacity in Real-World Applications of Public Sector Machine Learning', in *Presented at the 4th Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT/ML 2017), Halifax, Nova Scotia, Canada, 2017* (2017) (<https://arxiv.org/abs/1706.09249>).
- Veale M and Binns R, 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data' (2017) 4(2) *Big Data & Society* DOI: 10/gdcfnz.
- Veale M, Binns R and Ausloos J, 'When data protection by design and data subject rights clash' (2018) 8(2) *International Data Privacy Law* 105 DOI: 10/gdxthh.
- Veale M, Binns R and Edwards L, 'Algorithms That Remember: Model Inversion Attacks and Data Protection Law' (2018) 376 *Phil. Trans. R. Soc. A* 20180083 DOI: 10/gfc63m.
- Veale M, Binns R and Van Kleek M, 'Some HCI Priorities for GDPR-Compliant Machine Learning' in *The General Data Protection Regulation: An Opportunity for the CHI Community? (CHI-GDPR 2018). Workshop at ACM CHI'18, 22 April 2018, Montréal, Canada* (2018) (<https://arxiv.org/abs/1803.06174>).
- Veale M and Brass I, 'Administration by Algorithm? Public Management meets Public Sector Machine Learning' in K Yeung and M Lodge (eds), *Algorithmic Regulation* (Oxford University Press 2019).
- Veale M and Edwards L, 'Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling' (2017) 34(2) *Comput. Law & Secur. Rev.* 398 DOI: 10/gdhrtm.
- 'Better seen but not (over)heard? Automatic lipreading systems and privacy in public spaces' [2018] Presented at PLSC EU 2018.
- Veale M, Edwards L, Eyers D, Henderson T, Millard C and Staudt Lerner B, 'Automating Data Rights' in D Eyers, C Millard, M Seltzer and J Singh (eds), *Towards Accountable*

Bibliography

- Systems (Dagstuhl Seminar 18181)* (Dagstuhl Reports 8(4), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik 2018) DOI: 10/gffngz.
- Veale M, Van Kleek M and Binns R, 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'18)* (ACM 2018) DOI: 10/ct4s.
- Vedder A, 'KDD: The challenge to individualism' (1999) 1(4) *Ethics Inf. Technol.* 275.
- Venkatadri G, Lucherini E, Sapiezynski P and Mislove A, 'Investigating Sources of PII Used in Facebook's Targeted Advertising' [2018] *Proceedings on Privacy Enhancing Technologies*.
- Verwer S and Calders T, 'Introducing positive discrimination in predictive models' in B Custers, T Calders, B Schermer and T Zarsky (eds), *Discrimination and privacy in the information society* (Springer 2013).
- Vincent J, 'Can Deep Learning Help Solve Lip Reading?' (*The Verge*, 7th November 2017) (<https://www.theverge.com/2016/11/7/13551210/ai-deep-learning-lip-reading-accuracy-oxford>).
- von Schomberg R, *From the ethics of technology towards an ethics of knowledge policy & knowledge assessment* (European Commission 2007) (<https://perma.cc/Q7KQ-CQLE>).
- Vrandečić D and Krötzsch M, 'Wikidata: A Free Collaborative Knowledgebase' (2014) 57(10) *Commun. ACM* 78.
- Wachter S, 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 DOI: 10/gfc7bb.
- Wachter S, Mittelstadt B and Russell C, 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR' [2018] *Harv. J.L. & Tech.*
- Wagner B, 'Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping?' in E Bayamlioglu, I Baraliuc, LAW Janssens and M Hildebrandt (eds), *BEING PROFILED: COGITAS ERGO SUM* (Amsterdam University Press 2018).
- Weber M, 'Bureaucracy' in Gerth HH and Wright Mills C (trs), *From Max Weber* (Routledge 1958).
- Weiss K, Khoshgoftaar TM and Wang D, 'A Survey of Transfer Learning' (2016) 3(1) *Journal of Big Data* DOI: 10/gfkr2w.
- Wetenschappelijke Raad voor het Regeringsbeleid, *Big Data in een vrije en veilige samenleving (WRR-Rapport 95)* (WRR 2016) (<http://www.wrr.nl/publicaties/publicatie/article/big-data-in-een-vrije-en-veilige-samenleving/>).
- Wick MR and Thompson WB, 'Reconstructive expert system explanation' (1992) 54(1-2) *Artificial Intelligence* 33 DOI: 10/d529gf.

-
- Wickham H, 'Tidy Data' (2014) 59(1) *Journal of Statistical Software* 1 DOI: 10/gdm3p7.
- Wilcocks LP and Lacity MC, *Service automation* (Steve Brookes Publishing 2016).
- Wiles P, *Annual Report 2016: Commissioner for the Retention and use of Biometric Material* (Office of the Biometrics Commissioner 2017).
- Willenborg L and Waal T de, *Elements of Statistical Disclosure Control* (Springer 2012).
- Williams BO and Nagel T, 'Moral Luck' (1976) 50 *Proceedings of the Aristotelian Society*, Supplementary Volumes 115 DOI: 10/gfphtz.
- Winkler WE, 'Re-Identification Methods for Masked Microdata' in J Domingo-Ferrer and V Torra (eds), *Privacy in Statistical Databases* (Lecture Notes in Computer Science, Springer Berlin Heidelberg 2004) DOI: 10/d2jx8t.
- Woollard M, 'Administrative Data: Problems and Benefits. A perspective from the United Kingdom' in A Duşa, D Nelle, G Stock and GG Wagner (eds), *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences* (SCIVERO Verlag 2014).
- Yadav N, Yadav A and Kumar M, *An Introduction to Neural Network Methods for Differential Equations* (Springer 2015) DOI: 10/gfgrb3.
- Yamada T, Gohshi S and Echizen I, 'Use of Invisible Noise Signals to Prevent Privacy Invasion Through Face Recognition from Camera Images' in *Proceedings of the 20th ACM International Conference on Multimedia* (MM '12, ACM 2012) DOI: 10/gfksck.
- Yang Q, Zimmerman J, Steinfeld A, Carey L and Antaki JF, 'Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help' in *Proceedings of the 2016 SIGCHI Conference on Human Factors in Computing Systems (CHI'16)* (ACM 2016) DOI: 10/gddkjt.
- Ye L and Johnson P, 'The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice' (1995) 19(2) *MIS Quarterly* 157 DOI: 10/brgtbj.
- Yeom S, Giacomelli I, Fredrikson M and Jha S, 'Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting' [2018] *IEEE Computer Security Foundations Symposium (CSF 2018)*.
- Yeung K, 'Hypernudge': Big Data as a Mode of Regulation by Design' (2017) 20(1) *Information, Communication & Society* 118 DOI: 10/gddv9j.
- Zafar MB, Valera I, Gomez Rodriguez M and Gummadi KP, 'Fairness beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment' in *Proceedings of the 26th International Conference on World Wide Web* (International World Wide Web Conferences Steering Committee 2017) DOI: 10/gfgq8r.
- Zaret E and Meeks BN, 'Kozmo's digital dividing lines' [2000] MSNBC.

- Zarsky T, 'Transparency in data mining: From theory to practice' in B Custers, T Calders, B Schermer and T Zarsky (eds), *Discrimination and privacy in the information society* (Springer 2013).
- 'Incompatible: The GDPR in the Age of Big Data' (2017) 47 Seton Hall L. Rev. 995.
- Zelevnikov J, 'The Split-Up project: Induction, context and knowledge discovery in law' (2004) 3 Law, Probability & Risk 147 DOI: 10/bt4rcx.
- Zeng J, Ustun B and Rudin C, 'Interpretable classification models for recidivism prediction' (2017) 180(3) Journal of the Royal Statistical Society: Series A (Statistics in Society) 689 DOI: 10.1111/rssa.12227.
- Zhou Z.-H, 'Learnware: on the future of machine learning' (2016) 10(4) Front. Comput. Sci 589 DOI: 10/cwdj.
- Zhou Z, Zhao G, Hong X and Pietikäinen M, 'A Review of Recent Advances in Visual Speech Decoding' (2014) 32(9) Image and Vision Computing 590 DOI: 10/f6gqtq.
- Ziewitz M, 'Governing Algorithms: Myth, Mess, and Methods' (2016) 41(1) Science, Technology, & Human Values 3 DOI: 10/gddv9k.
- Žliobaitė I and Custers B, 'Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models' (2016) 24(2) Artif. Intel. & Law 183 DOI: 10/gfgt9b.
- Žliobaitė I, Kamiran F and Calders T, 'Handling Conditional Discrimination' in *2011 IEEE 11th International Conference on Data Mining* (IEEE 2011) DOI: 10/fzxwfp.
- Zuboff S, 'Big other: surveillance capitalism and the prospects of an information civilization' (2015) 30(1) Journal of Information Technology 75 DOI: 10/gddxpv.
- Zuiderveen Borgesius F, Kruikemeier S, Boerman SC and Helberger N, 'Tracking Walls, Take-It-Or-Leave-It Choices, the GDPR, and the ePrivacy Regulation' (2017) 3(3) European Data Protection Law Review 353 DOI: 10/gfsh4x.
- Zuiderveen Borgesius F and Poort J, 'Online Price Discrimination and EU Data Privacy Law' (2017) 40(3) Journal of Consumer Policy 347 DOI: 10/gdz28f.
- Zunger Y (*Twitter* [[@yonatanzungger](https://twitter.com/yonatanzungger)], 28th June 2015) (<https://perma.cc/7PMP-ZLT9>).