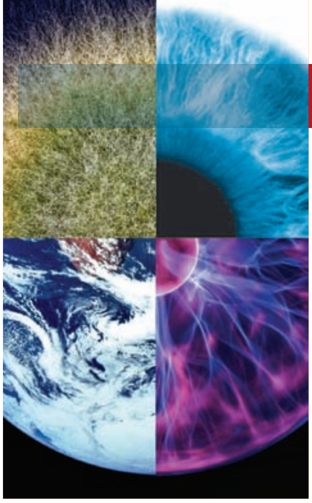

Editor: **Elena Simperl,** University of Southampton, e.simperl@soton.ac.uk

Data Mining and Automated Discrimination: A Mixed Legal/Technical Perspective

Laura Carmichael and Sophie Stalla-Bourdillon, *University of Southampton*
Steffen Staab, *University of Koblenz-Landau and University of Southampton*

Many industries are taking advantage of data analytics and a wealth of accessible (personal) data on the Web to bring about various automated, socially sensitive decisions. These decisions are socially sensitive in the sense that they are likely to have major impacts on the individuals subject to them (for example, being refused a mortgage).

The move toward more complex forms of socially sensitive decision making (for example, through machine learning) has various benefits, including the greater speed in which decisions are made, the ability to process multiple decisions at the same time, and the potential for an unprecedented amount of data to enrich the decision-making process. For instance, Kreditech Group (www.kreditech.com), a German technology company and financial services provider, openly uses machine learning for consumer credit scoring.[1] Its big data credit scoring algorithm processes about 20,000 constantly changing data points, including a person's online shopping and browsing behavior.

Despite the benefits of such automated decision making (for example, greater personalization and predictive analysis), how do we as IT professionals know whether such automated socially sensitive decisions are fair and fit for purpose? In other words, how can we be certain that that the use of data analytics for socially sensitive decision making does not lead to unjust discrimination against certain groups of individuals because of their age, ethnicity, gender, or sexual orientation?

Automated Discrimination

Automated discrimination refers to instances in which unfair treatment of certain groups of individuals occurs as a direct result of data analytics. For instance, it can manifest as unfair pricing and marketing of products and services.[2] This is also known as "weblining."[3]

Although most providers do not want their algorithms to be the source of discrimination, automated discrimination could occur unintentionally and through proxies. A proxy is defined as a nonprotected attribute used in lieu of a protected characteristic; for example, a neighborhood could be used as a proxy for ethnicity. For instance, in the 1980s, a computer program designed to screen applicants to St. George's Hospital Medical School was found to have automatically (although unknowingly) discriminated against female and minority applicants.[4] The key problem is that data-mining algorithms could inherit (historic) biases and prejudices; the patterns they uncover can merely reflect current inequalities.[5]

Discovery and Prevention

The discovery and prevention of automated discrimination is not without difficulty. Although data-mining algorithms are well-known and understood, their actual results (that is, classifying an individual or predicting its net value for the company—for example, as realized in a neural network) might be too intricate even for technical experts to understand. In theory, we could scrutinize every aspect of such a predictor, but in practice, we often would not understand its internal workings. Thus, we often are bound to assess only the (un)fairness of its treatments from how it behaves with regard to actual individuals.

Furthermore, it is important that any allegation of automated discrimination is thoroughly scrutinized. In most cases, a simple statistical check is not sufficient. For instance, a trend in a number of datasets

1541-1672/16/$33.00 © 2016 IEEE
Published by the IEEE Computer Society

can be reversed when the datasets are combined—that is, Simpson's paradox.

For an example of Simpson's paradox, consider efforts by P.J. Bickel and colleagues to analyze whether the decision to select applicants to the University of California, Berkeley, in 1973 was influenced by gender.[6] Bickel and colleagues started with the "simplest approach"—to aggregate the data for the entire campus. In total, about 44 percent of the male applicants were admitted to the university, whereas about 35 percent of the female applicants were admitted.[6] At first glance, it appeared there had been some form of discrimination based on gender—however, this was later found to be misleading. The initial statistical analysis did not consider how the entry process to different departments varied. Some departments attracted fewer female applicants than others—for example, only 2 percent of applicants to mechanical engineering were female, but overall acceptance rates to mechanical engineering were higher than for other departments.[6] Disaggregating the data, the researchers examined the data of each department individually. In fact, the majority of departments showed a "small but statistically significant bias" in favor of female applicants.[6]

State of the Art: Discrimination Discovery

Since 2008,[7] the data-mining community has directly responded to the challenges posed by automated discrimination through the emerging area of discrimination-aware data mining (DADM). DADM is focused on the discovery of unfair discriminatory practices and outcomes, which are concealed within datasets of historical decisions.[8,9] A principal DADM approach centers on the extraction and analysis of discriminatory classification rules.[7]

A significant proportion of the literature highlights the importance of "legally grounded rules" and "legally protected groups" as part of DADM. For instance, DADM research already recognizes a need to use a legal definition of discrimination.[9] Despite the recognition of legal factors, there does not appear to be a proposed DADM model, tool, or framework that sufficiently addresses the complexities of a specific legal framework (for example, the UK) and its particular jurisdictional constraints.

Hence, the AI community cannot confront the challenges posed by automated discrimination in isolation. A greater legal understanding is therefore crucial to enrich the existing DADM research and ensure that those directly responsible for socially sensitive decision-making algorithms remain legally compliant. There is therefore an opportunity for technical and legal experts to come together and address the challenges of automated discrimination discovery and prevention.

Interdisciplinarity: Unlawful Discrimination

The legal framework for automated discrimination is difficult for IT professionals to navigate. It is not only multijurisdictional, but it also spans various legal areas, including equality and data protection laws. For instance, two recent US reports highlight how automated discrimination is being dealt with under the US legal framework.[10,11] Notably, Article 14 of the European Convention on Human Rights—the reach of which extends far beyond the 28 member states of the EU and remains relevant even after Brexit—prohibits discrimination.

While Article 21 of the Charter of Fundamental Rights of the EU also prohibits discrimination, equality law continues to be largely regulated on a national scale in the EU. For instance, in the UK, Section 4 of the Equality Act of 2010 provides nine categories of protected characteristics: age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation. However, this list is restrictive, because data-mining algorithms could use other sensitive attributes beyond this list that could cause unfair treatment to certain groups of individuals. For instance, there is no category protecting socioeconomic status (although France, for example, has recently added poverty as a protected characteristic[12]). Moreover, pursuant to the UK Equality Act of 2010, data-mining algorithms could indirectly discriminate where they use proxies (background data such as neighborhood) for particular legally protected characteristics (for example, ethnicity).

As we mentioned earlier, equality law is only one aspect of the legal framework for automated discrimination. The recently adopted European General Data Protection Regulation (GDPR) 2016/679 provides a unified approach across the 28 member states of the EU. Similar to the protected characteristics outlined by the Equality Act, Article 9 of the GDPR outlines several "special categories of personal data" whose processing is prohibited—subject to 10 exceptions. Note that Article 9 includes some categories of data that are not listed as protected characteristics by Section 4 of the Equality Act—those are biometric and genetic data. However, it does not explicitly mention some of these protected characteristics, such as age, gender reassignment, and marriage and civil partnership.

Furthermore, pursuant to Articles 13(2)(f) and 14(2)(g) of the GDPR, data subjects have the right to be informed about the logic involved in automated decision making. However, where data-mining algorithms are concerned, will data controllers be able to fully explain this logic? Although technical methods are available (for example, Erik Štrumbelj and Igor Kononenko examine a sensitivity analysis-based method for explaining prediction models[13]), they mostly say, "This feature"—for example, income—"weighs *x*, and as a result you have been assigned to tariff-A as opposed to tariff-B." This statement is neither logically crisp nor completely true. The statement suggests a linear regression, but most data-mining algorithms do not constitute linear regression. Thus, this explanation is only an approximation of what is going on underneath. Furthermore, can data subjects be certain that such logical explanations will be expressed in terms that the layperson understands?

Finally, it is unclear to what extent personal data reused by decision makers must be accurate. The data subject has a right to rectification under Article 16 of the GDPR—that is, the data controller must remedy inaccurate data about that individual without delay. However, in the recent past, different legal approaches have been taken to data accuracy. In the case of *Smeaton v. Equifax*,[14] the UK's Court of Appeal held that personal data reused by credit reference agencies did not have to be absolutely accurate under the UK Data Protection Act 1998, although agencies should take reasonable steps to ensure that it is up to date. However, the landmark "right to be forgotten" ruling (taken by the Court of Justice of the European Union in the case of *Google Spain v. AEPD*,[15] which centered on auction notice of a repossessed home) gave rise to a duty to process accurate and timely data. Thus, it will be interesting to see how Article 16 is interpreted in the coming years, and how the legal position on personal data accuracy unfolds across sectors.

Accountability and Transparency

To reiterate, because the legal framework is multifaceted and difficult to navigate, a dialogue between computer scientists and lawyers is therefore crucial for the development of a robust legal–technical approach to the prevention and discovery of automated discrimination. A key challenge for policy makers and regulators is how data-mining algorithms can be made more accountable, both legally and technically, to the people they are profiling. Greater transparency could be achieved by placing a limited duty of disclosure on those responsible for such automated decision making. This may involve the release of (redacted) de-identified input and output data and a discrimination impact assessment, in addition to data-protection obligations (in particular, Article 35 of the GDPR). However, at the same time, it must be considered whether it is possible to formulate such a duty without jeopardizing intellectual property rights. The French Digital Republic Bill of 2016, for example, requires public sector bodies of more than 50 employees to make publicly available, in an open and easily reusable format, the rules defining the main algorithmic processing used in the accomplishment of their missions when such processing forms the basis of individual decisions.[16]

Furthermore, the capture of robust provenance information that covers organizational practices, processes, and principles pertaining to automated discrimination will be vital, not only for transparency and accountability but to ensure data accuracy and uphold the right to rectification.

Challenges for AI and the Law

Although DADM offers an excellent body of knowledge to build on, several challenges remain for the AI and legal communities to jointly confront. The principal challenge is the development of interdisciplinary tools that allow for the targeted use of data analytics and data mining to uncover discrimination and sufficiently address the diversity of enforcement-related issues. Greater interdisciplinary understanding is also required of the relationship between existing sociocultural, legal, and technical safeguards that aim to minimize unfair treatment.

Therefore, there is a pressing need for a legal–technical argumentation framework that helps decision makers assess the fairness of their black-box decision-making systems by providing arguments for and against allegations of automated discrimination. This framework must draw together existing machine learning algorithms that discover and prevent unfair treatment (such as DADM) as well as consider legal norm compliance, which spans a wide range of pertinent legal measures (for example, equality, data protection, and consumer protection laws).

As part of the development of this framework, it would be useful to examine the effectiveness of different quantitative and qualitative methods. We can do this in several ways. First, look for correlations within all input data: What input and output data correlate with legally protected characteristics and special categories of personal data? Is there potential for proxies?

Second, consider the "comparative individual" counterargument. For instance, in the context of automated decision making within the insurance industry, are people with the same or very similar risk category scores placed on the same tariffs? If men and women of the same profession with the same risk assessment pay the same, there is no discrimination. However, one counterexample is probably not enough to disprove discrimination.

Third, apply the test of reasonableness: Is there a justification for this discrimination? For instance, under Section 13 of the Equality Act of 2010, age discrimination can be justified on the grounds that the discrimination is for the purposes of a legitimate aim. An example of a legitimate aim might be where an applicant for a firefighter job is asked to undergo a fitness test—it is more likely that a younger person will pass.[17] However, legitimate aims are assessed on a case-by-case basis.

Fourth, assess internal processes and procedures—for instance, to what extent are those directly responsible for a data-mining algorithm transparent and accountable? Has there been a discrimination impact assessment? Is there a transparency report?[18] What provenance information is available? This latter question is particularly important because the data-mining process might be working well, but a "wrong" selection of data in the overall decision-making process could lead to problematic results—even if the test of the data mining would not show negative effects.

Finally, use rule-discovery algorithms to discover rules within the data-mining algorithm. However, remember that rule-discovery algorithms have limitations—that is, multiple rules can give the same or similar approximations with different discriminatory outcomes.

Promoting Fair Treatment

Automated discrimination is just one of many issues that must be addressed in the overall effort to promote fair treatment. Although working toward the legal–technical argumentation framework briefly outlined in this article is no panacea, it could potentially help those directly involved with the design, development, and use of data analytics to better safeguard data subjects from discrimination. Furthermore, greater accountability and transparency could better inform data subjects about how our digital footprints are (mis)used and about our associated rights (for example, under equality and data protection laws).

As the digital age matures, we become more connected (for example, through the Internet of Things), our digital footprints continue to expand, and more socially sensitive decisions are generated through data analytics. Automated discrimination only has the potential to increase. We need to recognize that advanced forms of automated socially sensitive decision-making systems have the potential to discriminate just as their nonautomated counterparts have done in the past.[4] ∎

Acknowledgments

We thank the Web Science Institute (WSI) for supporting Part I of our research project—Observing and Recommending from a Social Web with Biases—as part of its Pump-Priming Projects 2015–2016, as well as Southampton Law School for supporting Part II. For more information about this research, you can find our preliminary working report at http://eprints.soton.ac.uk/id/eprint/393352.

References

1. E. Reynolds, "The Next Industrial Revolution Is Coming—And It Will Be Fuelled by AI," *Wired*, 2016, 22 June 2016; www.wired.co.uk/article/ai-revolution-alexander-graubner-muller-kreditech.
2. A. Danna and O.H. Gandy Jr., "All That Glitters Is Not Gold: Digging Beneath the Surface of Data Mining," *J. Business Ethics*, vol. 40, no. 4, 2002, pp. 373–386.
3. M. Stepanek, "Weblining: Companies Are Using Your Personal Data to Limit Your Choices—And Force You to Pay More for Products," *Bloomberg Business Week*, 3 Apr. 2000; www.bloomberg.com/news/articles/2000-04-02/weblining.
4. S. Lowry and G. Macpherson, "A Blot on the Profession," *British Medical J.*, vol. 296, no. 6623, 1988, pp. 657–658.
5. S. Barocas and A.D. Selbst, "Big Data's Disparate Impact," *California Law Rev.*, vol. 104, 2016, pp. 1–62.
6. P.J. Bickel, E.A. Hammel, and J.W. O'Connell, "Sex Bias in Graduate Admissions: Data from Berkeley," *Science*, vol. 187, no. 4175, 1975, pp. 398–404.
7. D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-Aware Data Mining," *Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2008, pp. 560–568.
8. S. Ruggieri, D. Pedreschi, and F. Turini, "DCUBE: Discrimination Discovery in Databases," *Proc. ACM SIGMOD Int'l Conf. Management of Data*, 2010, pp. 1127–1130.
9. A. Romei, S. Ruggieri, and F. Turini, "Discrimination Discovery in Scientific Project Evaluation: A Case Study," *Expert Systems with Applications*, vol. 40, no. 15, 2013, pp. 6064–6079.
10. *Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues*, Federal Trade Commission, Jan. 2016; www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.
11. C. Muñoz, M. Smith and D.J. Patil, *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, May 2016; www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.
12. *Loi n°2016-832 du 24 Juin 2016 Visant à Lutter contre la Discrimination à Raison de la Précarité Sociale*, 2016 (in French); www.legifrance.gouv.fr/eli/loi/2016/6/24/AFSX1514889L/jo/texte.
13. E. Štrumbelj and I. Kononenko, "Explaining Prediction Models and Individual Predictions," *Knowledge and*

Information Systems, vol. 41, no. 3, 2014, pp. 647–665.

14. *Smeaton v. Equifax Plc*, England and Wales Court of Appeal, Civ. 108, 2013.
15. *Google Spain v. AEPD*, EU European Court of Justice, C-131/12, 3 WLR 659, 2014.
16. *Projet de Loi pour une République Numérique: Procédure Accélérée Engagée par le Gouvernement le 9 Décembre 2015*, Article 4, item 802 (in French); www.senat.fr/dossier-legislatif/pjl15-325.html#timeline-7.
17. "Justifying Discrimination," *Citizens Advice*, 2016; www.citizensadvice.org.uk/discrimination/what-are-the-different-types-of-discrimination/justifying-discrimination.
18. A. Datta, S. Sen, and Y. Zick, "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems," *Proc. 37th IEEE Symp. Security and Privacy*, 2016, pp. 598–617.

Laura Carmichael is a research fellow at the University of Southampton. Contact her at laura14carmichael@gmail.com.

Sophie Stalla-Bourdillon is an associate professor in IT law and director of the Institute for Law and the Web at the University of Southampton. Contact her at s.stallabourdillon@soton.ac.uk.

Steffen Staab is a professor for database and information systems and head of the Institute for Web Science and Technologies at the University of Koblenz-Landau and holds a chair for Web and Computer Science at the University of Southampton. Contact him at s.r.staab@soton.ac.uk.