Economic Models of (Algorithmic) Discrimination

Bryce W. Goodman
Oxford Internet Institute
University of Oxford
bwgoodman@gmail.com

Abstract

Organizations ranging from consumer finance to criminal justice increasingly rely on data mining and algorithmic decision-making as a compliment to or, in some cases, substitute for human judgement. This transition is motivated by improved efficiency, accuracy and reduced cost. However, it also raises important concerns. Chief among them is algorithmic discrimination, which occurs when certain groups or individuals unfairly receive unfavorable treatment as a result of algorithmic decision-making. Different algorithms discriminate for different reasons; identifying and rectifying algorithmic discrimination requires attention to these key differences. To that end, this paper draws upon literature from economics to differentiate between various sources and consequences of algorithmic discrimination. In particular, I show how the economic theories of taste-based and statistical discrimination apply in an algorithmic setting. The contribution of the work is threefold: to bridge work on algorithmic discrimination in computer and social sciences, to develop generalizable mathematical models for different types of algorithmic discrimination, and to identify sources of algorithmic discrimination that are unaddressed in current literature.

1 Introduction

Algorithms, and the data they process, play an increasingly important role in decisions with significant consequences for human welfare (Mayer-Schonberger and Cukier, 2013). This includes the allocation of both public and private goods (credit, employment, welfare, etc.). It is thus crucial that algorithmic decision making is considered not only in terms of technical performance, but also in terms of (re)distributive effects.

To date, research on algorithmic discrimination has tended to fall in one of two camps. The first is located within the territory of computer engineering, and is primarily focused on the technical aspects of algorithmic discrimination (e.g. Cus, 2013; Hajian et al., 2012). The second falls within the social sciences and is generally focused on documenting, describing and theorizing about the ethical, social and legal aspects (ELSA) of algorithmic discrimination (e.g. Valentino-DeVries et al., 2012), (Richards and King, 2013).

While the first camp is sensitive to the technical nuances of algorithm design, it does not engage with the wider social context within which algorithms operate, or acknowledge important conceptual differences between various types of discrimination. Further, its exclusive focus on developing formal methods for discrimination detection and removal can make the work obscure to readers without a technical background. This is problematic insofar as algorithmic discrimination is an issue that cuts across not only computer science but also law, sociology, ethics, etc. At the same time, the second camp tends to be both more widely accessible and attuned to the nuances of social context, but it does not always distinguish between how different types of algorithmic discrimination may arise.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

To bridge these camps, we require an account of algorithmic discrimination that is sensitive to both technical and social dimensions. Fortunately, economics has an established tradition of using formal methods to describe social phenomena (e.g. Schelling, 1971; Ross, 1973; Grossman and Stiglitz, 1980).

Within economics, discrimination is commonly defined as the less favorable treatment of a certain group, defined by some arbitrary characteristic(s), e.g. race or gender, which are otherwise indistinguishable in terms of some economically relevant set of characteristics, e.g. productivity (Becker, 1957). For example, discrimination occurs when members of one group are paid less than members of another, even though both groups are equally productive.

In their survey of discrimination literature, Romei and Ruggieri 'present an annotated multidisciplinary biography' that transverses social, legal, economic and computer science literature. This paper draws upon literature from economics to develop formal models attuned to both the social and technical dimensions of algorithmic discrimination. It is structured around two models of discrimination discussed in economics:

1. Theories of taste-based or 'irrational' discrimination (Becker, 1957);
2. Theories of statistical or 'rational' discrimination (Aigner and Cain, 1977);

In each case, I develop a model by applying the relevant economic theory to discrimination in an algorithmic setting (i.e. where one or more economic agents is an algorithm). The resulting models provide a clear picture of how algorithmic discrimination manifests and persists under different assumptions and highlight aspects of algorithmic discrimination that have eluded previous literature.

2 Theories of taste-based or 'irrational' discrimination

One of the earliest models of discrimination in economics is from Becker's (1957) *The Economics of Discrimination*. Becker's "taste-based" model assumes that employers hold a "taste for discrimination", wherein they exhibit a preference for hiring members of a certain group a over another group b. In his original formulation, an employer's utility function is modified to include a "discrimination coefficient", which represents the employer's discriminatory preference. The result is that the discriminatory employer is only willing to hire a worker from group b if

$$w_a - w_b > d \tag{1}$$

where w_a and w_b are the wages for a and b and d is the coefficient of discrimination.

2.1 How algorithms acquire a taste for discrimination

An algorithm will acquire a taste for discrimination if the relationship between the predictor and target variables in the dataset reflects discriminatory treatment (Cus, 2013). For example, a discriminatory manager may give non-white employees N lower ratings compared to whites W even though they are equally productive.[1] In this case, the rating Y may be given by the following formula:

$$Y = XB - aZ + e \tag{2}$$

where B and a are positive coefficients, X is a measure of actual productivity characteristics, $Z = 1$ indicates employee is non-white, and e is an unbiased error term.[2]

When a model is trained on the resulting dataset, the target variable Y' is the job rating by the (discriminatory) manager, while the predictor variables are some measure of performance characteristics X and a variable indicating group membership Z.

The accuracy of the model will be 100% if

$$Y' = Y = XB - az + e \tag{3}$$

If we assume that both groups are equally productive $X_N = X_W \equiv X$ and that e is independent from group membership $e_N = e_W \equiv e$, then

$$Y_N = XB - a + e < XB + e = Y_W \tag{4}$$

[1] In this example, we assume that the only thing that ought to count in a job performance rating is productivity.

[2] If we wish to model discrimination on the basis of multiple characteristics, we can simply consider Z a binary vector of group memberships and a a vector of associated coefficients of discrimination

Therefore in the perfect model (where accuracy = 100%), a nonwhite applicant will, *ceteris paribus*, receive a lower prediction than a white applicant.

2.2 Limits of the taste-based account

One consequence of the taste-based account is that firms employing discriminatory algorithmic profiling for hiring decisions may be at a long-term economic disadvantage.[3] This is because predictions are based on variables (e.g. race) that are not actually pertinent to productivity. Thus, on the taste-based account, we can rely on free markets, not policy, to end discrimination.

Unfortunately, things are not so simple. The taste-based account makes two key assumptions:

1. Group membership is independent from productivity
2. Employers explicitly prefer particular groups

In many cases, however, both assumptions are false. Consequently, a different account is needed.

3 Theories of statistical or 'rational' discrimination

An alternative to the taste-based account is provided by so-called 'statistical models' of discrimination proposed by Phelps (Phelps, 1972), Arrow (Arrow, 1973) and Aigner and Cain (Aigner and Cain, 1977). These models are based upon the idea that firms have limited information about applicants, and so are likely to utilize easily observable traits such as gender, race, education, etc. as proxies for productive characteristics. This effectively re-presents discrimination as "the solution to a signal extraction problem" (Autor, 2003).

Statistical discrimination differs from taste-based discrimination in that the decision maker is not intrinsically averse to any particular group per se. Rather, discrimination in this context is the 'rational' response to a situation where available information is limited and new information is hard to come by. Discrimination takes place when a decision-maker basis her individual predictions on group averages, or treats signals (e.g. test scores) from one group as more reliable than the other (Aigner and Cain, 1977).

3.1 How algorithms rationalize discrimination

In automated profiling, statistical discrimination can occur if a naive Bayesian method is employed and the difference in class probabilities for different demographic groups are sufficiently large. Consider a case where an algorithm is used to predict the future recidivism of an inmate up for parole. Let X be a vector of probabilities for recidivism based on behavioral factors (e.g. whether the inmate is a repeat offender, prison behavior, severity of crime, etc.) and Z be a vector of probabilities based on demographic factors (e.g. race, socio-economic class, etc.). The estimated risk of recidivism $E(Y)$ may be given by the formula:

$$E(Y|X,Z) = \frac{P(X|Y) * P(Z|Y) * P(Y)}{P(X) * P(Z)} \tag{5}$$

If the conditional probability of recidivism is greater for one demographic than another, $P(Z_a|Y) > P(Z_b|Y)$, a case could arise where an individual from demographic Z_a is denied parole but an individual from Z_b is released, even though they have the same behavioral factors ($X_a = X_b \equiv X$, see fig. 1).

More troublingly, if the conditional probability of recidivism is much greater for Z_a than Z_b, the inmate from group a may be denied parole even though b has a higher probability based on behavioral factors (see fig. 2). In other words, the algorithm might assign a higher recidivism risk to an individual from group a than to someone from group b even though the member of group a exhibits "better behavior".

[3] This assumes other production costs are equal and that the market is competitive (Becker, 1957). Also known as the "dynamic implication" (Romei and Ruggieri, 2013).

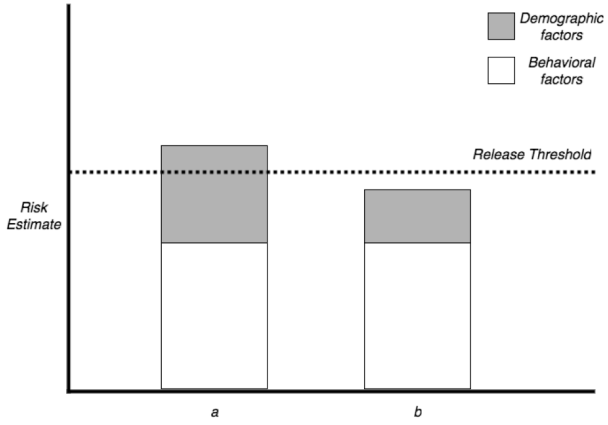

Figure 1: Risk Estimate: Same Behavioral, Different Demographic Factors

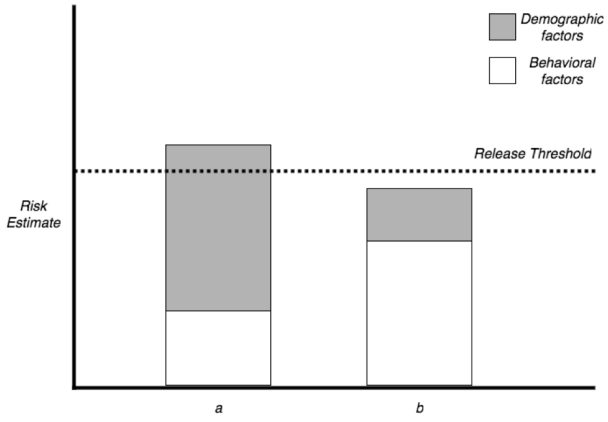

Figure 2: Risk Estimate: Different Behavioral and Different Demographic Factors

This case is particularly problematic if we believe that an inmate's behavior should play a greater role in determining whether she is released because it is, unlike demographic characteristics, within her control to change (Kahlenberg, 1996). More generally, we may believe that, in a particular situation, certain variables should carry a greater weight in determining the outcome *a priori*, that is, irrespective of whether they are in fact more predictive than some other set of variables. Recall: from a machine learning standpoint, all variables are equal, in the sense that there is no *a priori* preference for a model that predicts on the basis of an agent's intentional behavior versus her unchosen and pre-ordained demographic characteristics. This clashes with many normative theories of justice and equality, which hold that people are responsible first and foremost for those aspects of their life that are not due to chance. Thus Cohen writes that the "primary egalitarian impulse is to extinguish the influence on distribution" from factors that "[are] not the result of a gamble or risk which he could have avoided" (Cohen, 1989).

3.2 Uncertainty bias and the risk of risk aversion

The issue of under-representation in big data has been widely discussed (e.g. boyd and Crawford, 2012; Barocas and Selbst, 2014). However, there have so far been no attempts to formally characterize how it contributes to algorithmic discrimination. This section seeks to address this gap. In "Statistical Theories of Discrimination in Labor Markets", Aigner and Cain describe how risk aversion[4] can lead employers to discriminate if uncertainty, i.e. confidence interval associated with predictions, is unequal between groups (Aigner and Cain, 1977). In this section I argue that the same may be true for

[4]For the canonical account of risk aversion, see (Kahneman and Tversky, 1979)

algorithms that quantify, and seek to minimize, uncertainty associated with predictions: an *uncertainty bias* arises when under-representation and risk aversion leads to algorithmic discrimination.

Computational models used in machine learning typically rely on the assumption that samples are representative of the larger population and that the population characteristics measured by the sample will stay the same when the model is applied. If sampling is non-random, however, some groups may be over, under or un-represented in the sample. In the case of automated credit assessment, this would occur if firms use historic loan payment data to infer default risk, and some demographic groups received fewer loans compared to others. Creditors have less information about those prospective loan applicants. If the algorithm seeks to minimize uncertainty in its predictions, this could negatively impact the historically under-represented group's ability to access credit.

To illustrate, consider a case where group a and b have identical distribution of creditworthiness Y such that

$$Y_a, Y_b \sim N(a, \sigma^2) \tag{6}$$

but, historically, members of group a had less access to credit and are thus underrepresented in the training data ($n_a < n_b$). Let us also assume that an algorithm is risk averse, and will, *ceteris paribus*, prefer to lend when its predictions are more likely to be accurate (i.e. smaller confidence interval). The algorithm's risk-adjusted prediction Y^R is a function of two variables:

1. the predicted creditworthiness $E(Y)$ and
2. the *prediction uncertainty risk* R, which corresponds to the size of the confidence interval for that prediction

Such that

$$Y^R = E(Y) - R \tag{7}$$

Thus if we assume the same sample standard deviation in both groups, $\sigma_a = \sigma_b$, then the prediction uncertainty risk for group a, R_a, is increased relative to others

$$R_a = \frac{\sigma_a}{\sqrt{n_a}} > \frac{\sigma_a}{\sqrt{n_a}} = R_b \tag{8}$$

Therefore, if two members from group a and b each have the same creditworthiness estimate, $E(Y_a) = E(Y_b) \equiv E(Y)$, the algorithm will nonetheless treat the member of group a as having lower risk-adjusted creditworthiness since

$$Y_a^R = E(Y) - R_a < E(Y) - R_a = Y_b^R \tag{9}$$

This same phenomenon can be represented visually if we plot the distribution for two different groups and include error bars to indicate the associated confidence intervals (see fig. 3).

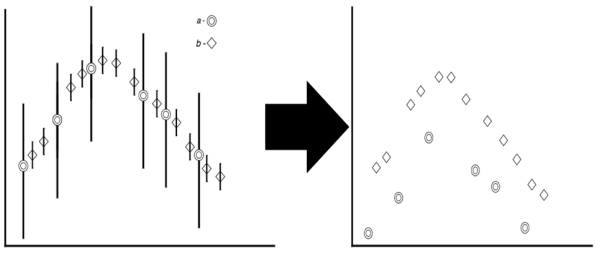

Figure 3: Risk Adjusted Predictions

The individuals from group a have a longer error bar, indicating that this group has higher prediction uncertainty risk. Note that individuals from both groups are part of the same normal distribution

$$\sim N(\mu_a, \sigma_a^2) = \sim N(\mu_b, \sigma_b^2) \tag{10}$$

The first graph in fig. 3 shows what happens if the classifier takes a risk-averse approach and scores members of each group at the lower end of the confidence interval. The predictions for group a have a larger confidence interval, and thus its members receive a much larger "uncertainty penalty" compared to members of group b.

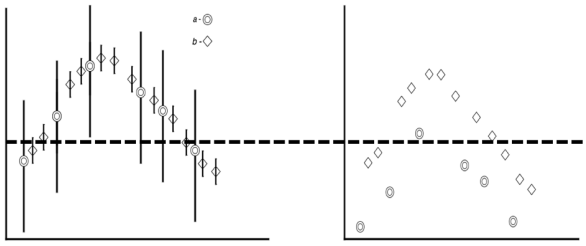

Figure 4: Risk Adjusted Predictions with Decision Boundary

Fig. 4 shows what happens when a decision boundary is introduced; notice two distinct phenomena. First, most members of group a fall below the threshold once their scores are adjusted to include prediction certainty. Second, some of the group a members below that threshold have an actual creditworthiness that is higher than for some of the members of group b that lie above the threshold. Risk aversion adds discriminatory uncertainty bias to an *ex ante* neutral algorithm by penalizing under-represented groups.

Before moving on, it is important to distinguish between the discrimination that is the result of known *uncertainty* about, and thus reticence to engage with, specific populations (discussed so far), versus discrimination that may arise from unrecognized discrepancies in an algorithm's *accuracy rate* across different populations. [5] In the first case, the algorithm exhibits an *uncertainty bias* in favor of particular populations, whereas in the latter the algorithm simply provides more or less accurate predictions for certain populations.

Both could, in practice, lead to discrimination: a bank may observe that the quality of predictions for a less represented group are inferior and thus be less likely to trust the algorithm's output. In this case, discrimination that arises could resemble the scenario described by Phelps 1972, where a signal is considered less reliable for a particular population. This leads the decision-maker to assign individuals from the less represented group a score that is closer to the population's hypothesized mean; in other words, to place a greater reliance on stereotypes. To the extent those stereotypes are inaccurate, discrimination occurs. However, if prediction errors for the under-represented group are randomly distributed, the algorithm may not exhibit any particular bias against the less represented group. With uncertainty bias, however, the algorithm *is* skewed against the under-represented group, since risk aversion will, *ceteris paribus*, lead the algorithm to penalize their scores.

3.3 Discrimination in Active Learning Algorithms

To date, all of the technical literature on algorithmic discrimination focuses on *batch learning*, where all data is collected prior to model training. This assumes that automated decision-making has no bearing on data generation for model training. In practice, however, this assumption is clearly false—the decisions made by algorithms deployed in the world today generate the data that will be used to train the predictive models of tomorrow. The issue of disparate representation is thus compounded in this more realistic case, where *active learning* algorithms adapt models to new data over time (see fig. 5).

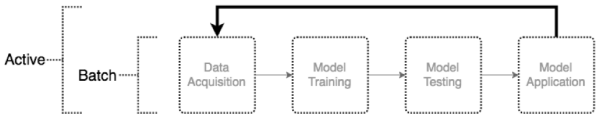

Figure 5: Batch vs. Active Learning Algorithms

[5] I am indebted to my reviewer for this observation.

In the case of the risk-averse credit-scoring algorithm, however, the "improvement" in performance depends on the relative strength of two competing factors. On the one hand, the a's above the threshold are, on average, much more creditworthy than the b's. This should improve the overall scoring of group a. On the other hand, this may be offset by the fact that there are now even fewer a's in the sample. Thus the prediction risk for group a will increase because there are fewer a's.

At issue is the fact that the algorithm does not merely discriminate in batch operations, but also increases discrimination by generating an increasingly unrepresentative training-set. More formally, the predicted creditworthiness Y at time t is also a function of the prediction risk R associated with that prediction

$$P(Y_t|R_t) = Y_t - R_t \tag{11}$$

where

$$R_t = \frac{\sigma_t^2}{\sqrt{n_t}} \tag{12}$$

The change in prediction risk between t and $t-1$ is simply

$$R_{t-1} - R_t = \frac{\sigma_{t-1}^2}{\sqrt{n_{t-1}}} - \frac{\sigma_t^2}{\sqrt{n_t}} \tag{13}$$

We have already shown that

$$n_{t-1} > n_t \tag{14}$$

So, assuming

$$\sigma_{t-1}^2 = \sigma_t^2 \equiv \sigma^2 \tag{15}$$

then

$$R_{t-1} - R_t = \frac{\sigma^2}{\sqrt{n_{t-1}} * \sqrt{n_t}} \tag{16}$$

3.4 When learners lag

Another real-world challenge not represented in the batch learning scenario concerns the rate at which the algorithm acquires information about different types of mistaken predictions. If all new data arrived at the same rate, an active learning algorithm would, at any point in time, be equally able to identify whether it had over-estimated (false positive) or under-estimated (false negative) an applicant's creditworthiness. It could use this data to update its model accordingly. This would allow an algorithm to correct for consistently under or over-estimating the creditworthiness of a particular group, and could thus serve as a safeguard against (irrational) discriminatory treatment. In practice, however, this is unlikely to always be the case.

In the current example, information will come more quickly for over-estimates than it will for under-estimates of credit-worthiness (Reichert et al., 1983). Consider an algorithm that decides on loans with a thirty-year period. If members of group a are predicted to have a $.05\%$ default risk over the duration of the loan but 20% default within the first five years, the algorithm can be corrected for overestimating creditworthiness. But what if the algorithm does the opposite, and mistakenly assumes a much higher level of default among members of group a compared to group b? The underestimation may be identified eventually, but unlike in the case of overestimation, there is no discrete event that can provide decisive evidence. Information about false positives comes in lumps, confirmation of true positives comes as a trickle, and identifying false negatives is impossible.[6] More formally, we can represent the predicted creditworthiness at time t as a function

$$Y_t = \gamma_1 Y_{t-n} - \gamma_2 O_t + \gamma_3 U_t \tag{17}$$

[6] In the credit-scoring industry, this issue is approached through "reject inference" (Crook and Banasik, 2004).

where Y_{t-n} are past predictions at time $t-n$, O_t is the number of known overestimates at time t (e.g. defaults), U_t is the number of known underestimates, and γ_n are weights. The formula states that data revealing past overestimates will cause the algorithm to decrease that group's credit score whereas evidence of past underestimates will have the opposite effect. If the algorithm's error is unbiased, overestimates and underestimates are equally likely. If information is equally available about both types of errors, the algorithm will self-correct.

However, in reality there is a delay between when an algorithm approves a loan and when the recipient demonstrates positive credit (e.g. on-time payments) or negative credit (e.g. late or missed payments). Consequently, the algorithm will not have access to the actual number of over-estimates or under-estimates, but only a portion

$$O'_t = (1 - L_O)O_t \tag{18}$$

$$U'_t = (1 - L_U)U_t \tag{19}$$

Where O_t and U_t are the actual over/underestimates at time t, L_n is a lag factor between 0 and 1, and O'_t and U'_t are the *known* over/underestimates at time t.

As discussed previously, data for underestimates have a greater *lag* compared to data for overestimates, such that $L_O > L_U$. Therefore, in the case where the algorithm's errors are unbiased and the number of initial over and underestimates are equal ($O_t = U_t$) the algorithms' data will contain more overestimates than underestimates. This results in a bias

$$\gamma_3(U_t - U'_t) - \gamma_2(O_t - O'_t) \tag{20}$$

or simply

$$\gamma_2 O'_t - \gamma_3 U'_t \tag{21}$$

The upshot is that it may be quicker for a group with historically good credit to get demoted (i.e. ranked as riskier) than for a group with historically poor credit to get 'upgraded' in an active learning algorithm. This feature highlights the fact that the initial distribution of observations in a dataset will have lasting consequences *even if* the algorithm's decision-rules are *prima facie* neutral. A similar observation underlies many of the arguments in favor of affirmative action, which calls for preferential treatment of historically disadvantaged groups as a means of ensuring distributive justice (Fullinwider, 2014). When it comes to "fair" algorithm design, a fuller discussion is merited, but outside the scope of this paper.

4 Conclusion

This paper has shown how economic theories of discrimination can usefully distinguish between different forms of algorithmic discrimination. By linking to economics, we are able to draw upon over 50 years of important work that considers the nuances of discrimination in a range of settings. The field of economics also offers an avenue for linking discourse on discrimination from the social and computer sciences through its tradition of generalizable mathematical models for social phenomena.

Of course this work is far from complete, and the scope of economic theories has been limited to the most "canonical" texts.[7] Further research should add to, challenge and refine these theories. In addition, this paper has only provided a very brief conceptual sketch of some of the sources of algorithmic discrimination that arise in an active learning environment. Much work is needed to provide a more complete taxonomy.

Finally, this paper has virtually ignored all of the proposed solutions for dealing with algorithmic discrimination. Much important work has been done in this area (e.g. (Datta et al., 2015; Cus, 2013; Hajian et al., 2012; Sandvig et al., 2014; Dive and Khedkar, 2014). However, so long as the problem of algorithmic discrimination is inchoate, solutions will be *ipso facto* incomplete.

References

Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases, volume 3 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-30486-6. URL `http://link.springer.com/10.1007/978-3-642-30487-3`.

[7] In particular, this paper has omitted a discussion of price discrimination (see Varian, 1985, 1989, 2006).

D. J. Aigner and G. G. Cain. Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review*, 30(2):175, Jan 1977. ISSN 00197939. doi: 10.2307/2522871.

K. Arrow. The theory of discrimination. *Discrimination in Labor Markets*, page 3–33, 1973.

D. Autor. Lecture note: The economics of discrimination i. Nov 2003. URL `https://pdfs.semanticscholar.org/8eb7/69b1a56161658cffd066f6039a1f4461e18d.pdf`.

S. Barocas and A. D. Selbst. Big data's disparate impact. *Available at SSRN 2477899*, 2014. URL `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2512208`.

G. Becker. *The economics of discrimination*. University of Chicago Press, Chicago, 1957.

d. boyd and K. Crawford. Critical questions for big data. *Information, Communication Society*, 15 (5):662–679, Jun 2012. ISSN 1369-118X. doi: 10.1080/1369118X.2012.678878.

G. A. Cohen. On the currency of egalitarian justice. *Ethics*, 99(4):906–944, 1989. ISSN 0014-1704.

J. Crook and J. Banasik. Does reject inference really improve the performance of application scoring models? *Journal of Banking Finance*, 28(4):857–874, 2004.

A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

R. Dive and A. Khedkar. An approach for discrimination prevention in data mining. *International Journal of Application or Innovation in Engineering and Management*, 3(6), Jun 2014. URL `http://www.ijaiem.org/Volume3Issue6/IJAIEM-2014-06-10-20.pdf`.

R. Fullinwider. *Affirmative Action*. Metaphysics Research Lab, Stanford University, winter 2014 edition, 2014. URL `http://plato.stanford.edu/archives/win2014/entries/affirmative-action/`.

S. J. Grossman and J. E. Stiglitz. On the impossibility of informationally efficient markets. *The American economic review*, 70(3):393–408, 1980.

S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer, and F. Giannotti. *Injecting discrimination and privacy awareness into pattern discovery*, page 360–369. IEEE, 2012. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6406463`.

R. D. Kahlenberg. *The remedy*. New York: Basic Books, 1996.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, page 263–291, 1979.

V. Mayer-Schonberger and K. Cukier. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. John Murray, Oct 2013. ISBN 978-1-84854-792-6.

E. S. Phelps. The statistical theory of racism and sexism. *The american economic review*, 62(4): 659–661, 1972.

A. K. Reichert, C.-C. Cho, and G. M. Wagner. An examination of the conceptual issues involved in developing credit-scoring models. *Journal of Business Economic Statistics*, 1(2):101–114, Apr 1983. ISSN 0735-0015. doi: 10.1080/07350015.1983.10509329.

N. M. Richards and J. H. King. Three paradoxes of big data. *Stanford Law Review*, 2013. URL `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2325537`.

A. Romei and S. Ruggieri. *Discrimination data analysis: a multi-disciplinary bibliography*, page 109–135. Springer, 2013. URL `http://link.springer.com/chapter/10.1007/978-3-642-30487-3_6`.

S. A. Ross. The economic theory of agency: The principal's problem. *The American Economic Review*, 63(2):134–139, 1973.

C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, 2014. URL `http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf`.

T. C. Schelling. Dynamic models of segregation†. *Journal of mathematical sociology*, 1(2):143–186, 1971.

J. Valentino-DeVries, J. Singer-Vine, and A. Soltani. Websites vary prices, deals based on users' information. *Wall Street Journal*, Dec 2012. ISSN 0099-9660. URL `http://www.wsj.com/articles/SB10001424127887323777204578189391813881534`.

H. R. Varian. Price discrimination and social welfare. *The American Economic Review*, 75(4): 870–875, 1985.

H. R. Varian. Price discrimination. *Handbook of industrial organization*, 1:597–654, 1989.

H. R. Varian. *Intermediate microeconomics*. WW Norton Company, seven edition, 2006.