Langenbucher, Katja

Working Paper

Consumer credit in the age of AI: Beyond anti-discrimination law

LawFin Working Paper, No. 42

Provided in Cooperation with:
Center for Advanced Studies on the Foundations of Law and Finance (LawFin), Goethe University

Suggested Citation: Langenbucher, Katja (2022) : Consumer credit in the age of AI: Beyond anti-discrimination law, LawFin Working Paper, No. 42, Goethe University, Center for Advanced Studies on the Foundations of Law and Finance (LawFin), Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/266418

Consumer Credit in The Age of AI – Beyond Anti-Discrimination Law

Law Working Paper N° 663/2022
November 2022

Katja Langenbucher
Goethe University Frankfurt, SciencesPo, Fordham Law School and ECGI

ECGI Working Paper Series in Law

Consumer Credit in The Age of AI – Beyond Anti-Discrimination Law

Working Paper N° 663/2022
November 2022

Katja Langenbucher

I am immensely grateful for having had the opportunity to present versions of this paper and receive invaluable feedback at: AI and democracy conference/SciencesPo; ECFR conference/University of Helsinki; Edinburgh University/FinTech Lecture; European Central Bank legal conference/Frankfurt; FinCoNet conference/OECD; FinTech conference/Luxembourg; Max-Planck-Institute;Bonn; Fintech conference/University of Hamburg; Fintech Symposion/University of Mannheim; Law and Society/Lisbon; NYU Law School/PRG group; PennLaw/Adhoc faculty workshop; RegHorizon conference/ETH Zürich;. I am especially grateful to Marion Fourcade, Talia Gillis, Sandy Mayson, Udo Milkau, Katherine Strandburg and Olivier Sylvain for critical comments and discussion. All remaining errors are mine.

Abstract

Search costs for lenders when evaluating potential borrowers are driven by the quality of the underwriting model and by access to data. Both have undergone radical change over the last years, due to the advent of big data and machine learning. For some, this holds the promise of inclusion and better access to finance. Invisible prime applicants perform better under AI than under traditional metrics. Broader data and more refined models help to detect them without triggering prohibitive costs. However, not all applicants profit to the same extent. Historic training data shape algorithms, biases distort results, and data as well as model quality are not always assured. Against this background, an intense debate over algorithmic discrimination has developed. This paper takes a first step towards developing principles of fair lending in the age of AI. It submits that there are fundamental difficulties in fitting algorithmic discrimination into the traditional regime of anti-discrimination laws. Received doctrine with its focus on causation is in many cases ill-equipped to deal with algorithmic decision-making under both, disparate treatment, and disparate impact doctrine. The paper concludes with a suggestion to reorient the discussion and with the attempt to outline contours of fair lending law in the age of AI.

Keywords: credit scoring methodology, AI enabled credit scoring, AI borrower classification, responsible lending, credit scoring regulation, financial privacy, statistical discrimination

JEL Classifications: C18, C32, K12, K23, K33, K40, J14, O31, O33

Katja Langenbucher
Professor
Goethe University Frankfurt
Theodor-W.-Adorno-Platz 3
60323 Frankfurt am Main, Germany
e-mail: langenbucher@jur.uni-frankfurt.de

LawFin Working Paper No. 42

Consumer Credit in the Age of AI – Beyond Anti-discrimination Law

Katja Langenbucher

DFG Center for Advanced Studies on the Foundations of Law and Finance | House of Finance | Goethe University
Theodor-W.-Adorno-Platz 3 | 60629 Frankfurt am Main | Germany
www.lawfin.uni-frankfurt.de

CONSUMER CREDIT IN THE AGE OF AI – BEYOND ANTI-DISCRIMINATION LAW

Katja Langenbucher[*]

ABSTRACT

Search costs for lenders when evaluating potential borrowers are driven by the quality of the underwriting model and by access to data. Both have undergone radical change over the last years, due to the advent of big data and machine learning. For some, this holds the promise of inclusion and better access to finance. Invisible prime applicants perform better under AI than under traditional metrics. Broader data and more refined models help to detect them without triggering prohibitive costs. However, not all applicants profit to the same extent. Historic training data shape algorithms, biases distort results, and data as well as model quality are not always assured. Against this background, an intense debate over algorithmic discrimination has developed. This paper takes a first step towards developing principles of fair lending in the age of AI. It submits that there are fundamental difficulties in fitting algorithmic discrimination into the traditional regime of anti-discrimination laws. Received doctrine with its focus on causation is in many cases ill-equipped to deal with algorithmic decision-making under both, disparate treatment, and disparate impact doctrine.[1] *The paper concludes with a suggestion to reorient the discussion and with the attempt to outline contours of fair lending law in the age of AI.*

TABLE OF CONTENTS

[*] Katja is a law professor at Goethe University, Frankfurt, Germany. She is also affiliated at SciencesPo, Paris, France, and regular visiting faculty at Fordham Law School, NYC, US. Katja is grateful to the DFG Project „Fair Scoring“/ LA1358/6-1. She also thanks the DFG Center for Advanced Studies on the Foundations of Law and Finance for their support (project FOR 2774). I am immensely grateful for having had the opportunity to present versions of this paper and receive invaluable feedback at: AI and democracy conference/SciencesPo; ECFR conference/University of Helsinki; Edinburgh University/FinTech Lecture; European Central Bank legal conference/Frankfurt; FinCoNet conference/OECD; FinTech conference/Luxembourg; Max-Planck-Institute;Bonn; Fintech conference/University of Hamburg; Fintech Symposion/University of Mannheim; Law and Society/Lisbon; NYU Law School/PRG group; PennLaw/Adhoc faculty workshop; RegHorizon conference/ETH Zürich;. I am especially grateful to Marion Fourcade, Talia Gillis, Sandy Mayson, Udo Milkau, Katherine Strandburg and Olivier Sylvain for critical comments and discussion. All remaining errors are mine.

[1] The paper uses the US terminology “disparate treatment” and “disparate impact” for a better reading experience. Much of what is said applies to EU law with its different terminology (“direct” and “indirect” discrimination). Throughout the paper, EU law is addressed explicitly where it differs from US law in crucial ways.

A. INTRODUCTION

The decision to hand out a loan and price interest rates includes an assessment of the borrower's credit risk. "Good" borrowers are separated from "bad" ones, credit default risk is predicted, and interest rates are calculated. Naturally, this involves distinguishing among applicants to make an informed choice. In the process, different groups of applicants emerge, some with excellent chances of obtaining attractively priced credit, others with reasonable chances, and some with low or no chances of affording interest rate payments or qualifying for a loan.

Faced with uncertainty about an applicant's credit default risk, with transaction costs and with imperfect competition lenders must reconstruct hidden fundamental information about borrowers. To do so, they rely on observable variables.[2] Historically, signals such as "capital, capacity, and character" were important clues towards fundamental information. Establishing them was a core part of the daily work of credit managers. Until the 1960s, a variable as qualitative and vague as "character" was "considered the foundation of consumer creditworthiness".[3] With advances in statistics, reasonably good forecasts could be established based on a limited list of input variables. When deciding which input variables to use, lenders faced a choice.[4] Comprehensive checks were slow in information gathering. They could find attractive candidates even with unusual attributes but were costly. Alternatively, a focus on limited and standardized input variables allowed for quick decisions which captured many, if not all cases. For most lenders, privileging speed and volume over comprehensive searches seemed, on balance, more attractive.[5] This led to enormous market expansion based on standardized decision-making criteria.[6] Politically, statistical approaches of this type were understood to replace "vague intuitions, personal prejudices, and arbitrary opinions".[7] They also held the promise of inclusion for groups which had found it hard to gain access to the financial system.

Today, we see a similar development with the advent of big data and machine learning algorithms.[8] Digital technology has achieved more efficient and lower-cost delivery of financial services than ever before.[9] Lenders can access data far beyond traditional financial variables, without compromising on speed and volume. One example is the borrower's cash flow and data with the lender. But there is much more to explore. Online payment history, performance on lending platforms, age or sex, job or college education, ZIP code, income or ethnic background can all be relevant to predict credit default risk. Depending on a jurisdiction's privacy laws more variables can be scrutinized. This includes, for instance, preferred shopping places, social media friends, political party affiliation, taste in music, number of typos in text messages, brand of smartphone, speed in clicking through a captcha exercise, daily work-out time, or performance in a psychometric assessment.

Empirically, countries such as the US have long faced imbalances in minorities' chances to access credit markets. Search costs for lenders are one of the reasons. An applicant with a low credit score might be a good credit risk and present an attractive business case for the lender. However, it is often not cost-efficient for the lender to invest in locating such "invisible prime"[10] applicants. The lender

---

[2] Bartlett et al. (2022); Brito/Hartley (1995); Parlour/Rajan (2001); Stiglitz/Weiss (1981); see Guseva/Rona-Tas (2001) on uncertainty and institutions which allow for reducing uncertainty to measurable risk.
[3] Lauer (2017), pp. 199 et seq. on the five variables used by the mail-order firm Spiegel in the 1930s; tracing the historical development: Citron/Pasquale (2014), pp. 8 et seq.
[4] Lauer (2017), p. 210.
[5] Lauer (2017), p. 210.
[6] Burrell/Fourcade (2021), p. 222 ("national trust infrastructure").
[7] Burrell/Fourcade (2021), p. 222.
[8] Burrell/Fourcade (2021), p. 222; Citron/Pasquale (2014), p. 4.
[9] FSB (2022), p. 11, worrying at the same time that some incumbent financial institutions prioritize market share through sales rather than operating profit.
[10] Term proposed by Di Maggio et al. (2021), p. 2.

will compare his search costs to the expected return on the loans he could hand out to applicants whose credit default risk is lower than what their score suggests. In the past, few lenders have found the expected return to be higher than the search costs.[11] As described in more detail below,[12] access to big data and ease of modelling via machine learning have significantly lowered search costs. For some invisible prime borrowers, this has raised hopes for inclusion through AI. At the same time, data on the extent of inclusion along those lines is still sparse and newer studies caution that minorities profit only disproportionately.[13]

Framed as a question of economic efficiency from the micro-perspective of the lender, there is nothing wrong with using as many variables as cost-efficient. Market forces are assumed to single out the meaningful signals, lowering costs of information for the lender. Signaling of this type is predicted to allow for risk-adjusted pricing according to the lender's business model. It is important to keep in mind that there is a broad spectrum of potential business models, ranging from risk-adjusted market-priced credit to finding opportunities for predatory lending.

Understood as a question of fair lending law, some signals must not be used. If a rejection in the context of a mortgage is motivated by race, color, religion, sex, disability, familial status or national origin, there might be liability under the US Fair Housing Act (FHA), Title VIII of the Civil Rights Act of 1968. The US Equal Credit Opportunities Act (ECOA) prohibits to deny a loan because of race, color, religion, sex, disability, marital status, age, national origin, receipt of income from a public assistance program or an applicant's good-faith exercise of any right under the Consumer Protection Act.

Under EU law, Directive 2004/113 of 13 December 2004 prohibits discrimination based on sex, including less favorable treatment of women for reasons of pregnancy and maternity. For a loan contract to qualify, it needs to be a "service which is available to the public irrespective of the person concerned" and "offered outside the area of private and family life".[14] Given the relevance of personal attributes, few contracts qualify. Additionally, the Directive highlights that it does not wish to "prejudice the individual's freedom to choose a contractual partner".[15] Outside the scope of this Directive, discriminatory lending practices have remained a matter of private law of the Member States.[16] The EU Consumer Credit Directive/2021,[17] for the first time, includes an explicit anti-discrimination provision as to consumers legally resident in the EU on grounds of nationality, place of residence, sex, race, color, ethnic or social origin, genetic features, language, religion, belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation.[18]

[11] Lauer (2017), p. 210.
[12] See below B.
[13] On Fintech more generally: FSB (2022), p. 12.
[14] Art. 3 para. 1.
[15] Art. 3 para. 2.
[16] For Germany see sec. 19, 20 *Allgemeines Gleichbehandlungsgesetz* (General Act on Equal Treatment).
[17] Proposal for a Directive on Consumer Credits of 30 June 2021, COM(2021)347 final. For better readability, I refer to this text as: EU Consumer Credit Directive/2021.
[18] Art. 6 of the proposal, see already recital (45) of the current Consumer Credit Directive 2008/48/EC.

Anti-discrimination laws are applicable irrespective of whether lenders feel that they are rejecting the applicant *because of* attributes which these laws protect. Lenders might wish to point out that they deny a loan because of high credit default risk, for which the protected attribute is merely a useful signal. However, as we will explore in more detail below, this is not an accepted defense in a disparate treatment case. It is prohibited for the lender to use an observable protected attribute, such as sex, to form beliefs about unobservable attributes, such as credit default risk.

Instead of relying explicitly on protected characteristics lenders might use proxies which correlate with a protected attribute. One example is part-time employment which often correlates with sex. More current examples concern big data, where first name, taste in music or preferred shopping place often correlate with race, sex, or religion.[19] As discussed in more detail below, the fact alone that a proxy correlates (even narrowly) with a protected attribute is not *per se* grounds for a case of direct discrimination/disparate treatment. Instead, these cases are where disparate impact/indirect discrimination doctrine comes into play. If the relevant proxy can be shown to trigger clusters of protected communities which are disproportionately faced with credit rejections or disadvantageous pricing, the proxy raises suspicion. Using it can violate anti-discrimination laws, even though the proxy is facially neutral and does not fall under the list of protected variables. There is no US Supreme Court guidance so far on whether disparate impact doctrine applies to retail credit decisions.[20] By contrast, the EU Directive mentioned above explicitly covers direct as well as indirect discrimination. Indirect discrimination is understood as an "apparently neutral provision, criterion or practice" which puts protected groups "at a particular disadvantage compared with other persons, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary".[21] Cases which qualify under US disparate impact doctrine will in most cases also qualify under EU indirect discrimination doctrine. However, even assuming that credit underwriting does, I describe below why it is rarely straightforward to establish a case.[22]

The rest of the paper is structured as follows: Section B summarizes some of the empirical studies on AI scoring and inclusion which have emerged over the last years. Section C offers a brief glance at economic theories on discrimination, before Section D reminds us of a number of computer scientist's concerns with historically biased training data and algorithmic fairness. In Section E the paper moves on to a detailed investigation of AI underwriting under the lens of US anti-discrimination law, including some remarks on EU law. The focus of this section is on the limits of received doctrine of disparate treatment and disparate impact doctrine. Section F suggests reorienting the discussion and offers a rough and preliminary outline of questions a future design will face.

[19] For econometricians trying to measure bias, this can lead to omitted-variable-bias if the relevant variable is not observed by the econometrician. The model will then attribute the effect of the missing variable to the observable variables which were included, Dobbie et al. (2019), p. 1.
[20] Langenbucher (2020) on comparing this to the EU approach on indirect discrimination; see Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 135 S.Ct. 2507 (2015) on applying disparate impact doctrine to the Fair Housing Act which prohibits discrimination in home lending.
[21] Art. 2 para. 2 lit. b Directive 2004/113.
[22] See below E.

The paper makes two main contributions to the debate. First, it brings out how received anti-discrimination law does not adequately capture discriminatory credit underwriting. A core reason for this is the traditional understanding of causation. Its conceptual underpinnings provide an ill fit with AI-based decision-making based on bundles of correlating variables which are often opaque even for the decider. Second, the paper suggests moving the discussion towards a regulatory design which is tailored to credit underwriting. It highlights four areas where policy work is required. These include setting a regulatory framework for quality control, adjusting current rules on credit reporting and financial privacy, rethinking transparency and responsible lending, and engaging in a debate on the costs of equal access.

B. GOOD INTENTIONS: INCLUSIONARY AI

Business models underlying credit pricing vary, depending on the availability and the costs of information. Scoring agencies deliver first signals, traditionally based on a short list of input factors. Many lenders refine this assessment by adding information of their own, using, for instance, cash flow and transaction data or proprietary algorithms to categorize potential borrowers. Empirically, US retail credit markets seem to move along highly standardized metrics.[23] The Consumer Financial Protection Bureau (CFPB) has found that input variables and modelling techniques are very similar across traditional lenders.[24] Multivariate linear regression analysis has been used to correlate a short list of (historically grown) variables (indicating past credit history and current credit usage) to consumer credit outcomes.[25] Compared with a sample of previous consumers with similar attributes, this is used to predict credit default risk.[26]

Applicants who are more costly to evaluate than the average often face obstacles in applying for a loan, even if they are low risk applicants.[27] Sometimes this is due to the scarcity of available information. In other cases the available information transmits a wrong or less meaningful picture of credit default risk.[28] For yet others, their future potential is not adequately reflected.[29] Such "thin-file applicants" or "credit invisible" have more recently been targeted by Fintech lenders who offer lower cost scoring.[30] Those who perform well under the novel combination of AI and big data embraced by Fintech scoring stand to gain if compared with standard metrics. This development has raised hopes of financial inclusion.[31]

---

[23] On a troubling symbiosis between traditional and payday lenders see Di Maggio et al. (2021).
[24] CFPB (2017), p. 6; on imperfect competition see Parlour/Rajan (2001).
[25] Aggarwal (2021), p. 46.
[26] CFPB (2017), p. 6
[27] On "error costs" in a different context but making a similar point: Hellman (2020), pp. 829, 836 et seq. on understanding error costs in their normative context.
[28] The CFPB estimates that 26 million Americans have no file with the major credit bureaus while another 19 million are unsociable because their credit file is either too thin or too stale to generate a reliable score, CFPB (2017), pp. 6–7.
[29] Di Maggio et al. (2021), p. 3.
[30] CFPB (2017), p. 15.
[31] Balyuk (2021); Bartlett et al. (2022), p. 55.

One important source of data are mobile phones. A study by Agarwal et al. (2021) on Indian markets lists data collected from the individual's mobile phone such as the mode of login, the apps installed, the number of calls, contacts and social connections as well as the kind of operating system.[32] When categorizing data, the authors work for instance with mobile phone data such as apps installed or operating systems, with data referring to the presence of social apps, preferred social networks or the number of contacts and calls, and, lastly, what the authors call "deep social footprint", understood as information obtained from call log patterns.[33] They provide many examples to illustrate their findings. Consumers which have no financial app installed increase their probability of default by 25 %. Having a dating app installed raises credit default risk by 17 %. Results hold after controlling for salary, age and education.[34] The authors find that a model that relies exclusively on mobile and social footprints outperforms traditional credit score models by 7 %.[35] They interpret this as a way to access "aspects of individuals' behavior which has implications for the likelihood of default"[36] – a thought not too far from last century's focus on "character".[37] Additionally, the authors use variables which measure frequency and duration of incoming, outgoing and missed calls as a way to capture an individual's "social capital". Again, they find that these measures are strongly correlated with the likelihood of default.[38]

Berg et al. (2018) analyze simple and easily accessible digital footprint variables using a data set of a German E-Commerce company. Their set contains only ten digital footprint variables. Among those are the device type, the operating system, the time of day of the purchase and a dummy for a typing error. The authors find, for example, that the difference in default rates between customers using an Apple and customers using an Android device is equivalent to the difference in default rates between a median credit score and the 80th percentile of the credit score.[39] The authors suggest that the variables they investigated provide a "proxy for income, character and reputation"[40] – again, referencing "character" to be inferred from digital data.

The online lending company Upstart provides another especially well-documented example for the promise of financial inclusion.[41] Upstart claims to outperform traditional scoring outfits not only as to all borrowers, but specifically as to those with traditionally low credit scores.[42] This covers approval decisions as well as interest rates.[43] Di Maggio et al. (2021)[44] find that "more than 30 % of borrowers with credit scores of less than 680 funded by Upstart over our sample period would have

[32] Agarwal et al. (2021), p. 4.
[33] Agarwal et al. (2021), p. 4.
[34] Agarwal et al. (2021), p. 5.
[35] Agarwal et al. (2021), p. 5.
[36] Agarwal et al. (2021), p. 5.
[37] Lauer (2017), p. 199.
[38] Agarwal et al. (2021), p. 6.
[39] Berg et al. (2018), p. 3.
[40] Berg et al. (2018), p. 3.
[41] Langenbucher/Corcoran (2022), p. 141; Di Maggio et al. (2021), p. 2 ("invisible primes").
[42] Di Maggio et al. (2021), p 3.
[43] Di Maggio et al. (2021), p. 4.
[44] Di Maggio et al. (2021), p. 4.

been rejected by the traditional model. We further find that this fraction declines as credit score increases, that is the mismatch between the traditional and the Upstart model is magnified among low-credit score borrower". The CFPB, which has investigated Upstart's business model, added that applicants with FICO scores from 620 to 660 were approved twice as frequently by Upstart if compared with a hypothetical lender using FICO. Applicants under the age of 25 were 32 % more likely to be approved and consumers with incomes under $50,000 were 13 % more likely to be approved.[45]

It is difficult to predict which variables will enhance an applicant's chances of being considered for a loan.[46] Education is one of the variables relied upon by Upstart which found that the better the educational background the higher the chances of profiting from Upstart's offer. By contrast, Agarwal et al. (2021), based on data from mobile phones and deep social footprints in India, found that the marginal benefits were likely to be higher for borrowers with low levels of education, indicating that relevant proxies and outcomes vary significantly across Fintech lenders and countries.[47]

There are more studies pointing towards financial inclusion through credit scoring based on alternative data. Agarwal et al. (2021) use proprietary data from a Fintech lender and find that mobile footprint, social footprint, and deep social footprints can expand credit access without adversely impacting the default outcomes. They suggest that 13 % of borrowers who were denied credit would be approved under the authors' alternate credit scoring model.[48] Bartlett et al. (2022) investigate the price of mortgages. They find that, today, Latinx and African American borrowers pay 7.9 and 3.6 basis points more in interest for home purchase and refinance mortgages because of discrimination. By contrast, Fintech algorithms discriminate 40 % less, with Latinx and African American borrowers paying 5.3 more in interest for purchase mortgages and 2.0 basis points for refinance mortgages originated on FinTech platforms.[49]

More cautiously, Fuster et al. (2022), using US data, show that a machine learning model is predicted to provide only slightly better access and only marginally reduce disparity in acceptance rates. They find more pronounced cross-group disparity leading to very different interest rates, a concern I explore further below.[50]

C. TASTE-BASED AND STATISTICAL DISCRIMINATION

Under the theoretical assumptions that the retail credit market is competitive, that lenders are risk-neutral and that they set interest rates contingent on borrowers' observable characteristics, we would

[45] CFPB (2019).
[46] Wachter (2022), p. 1 "algorithmic groups".
[47] Agarwal et al. (2021) using data from one of the largest Fintech lending firms in India.
[48] Agarwal et al. (2021), p. 8.
[49] Bartlett et al. (2022), pp. 31-32.
[50] Fuster et al. (2022), at p. 9 find the increase to be double the magnitude for Black and White Hispanic borrowers than for white non-Hispanic borrowers, see below E.III.1. on this point.

expect to be able to explain differences in access to credit and in pricing of loans with differences in credit risk.[51] Inequalities in access to loans and differences in pricing for minority groups could then be understood as "statistical discrimination": a population of loan applicants separated into groups according to their risk profile.[52]

Statistical discrimination is the natural result of an efficient evaluation of credit risk. By contrast, under these same assumptions, we would not expect a situation commonly described as "taste-based" discrimination. Taste-based discrimination occurs where lenders' individual preferences, such as a dislike of certain minority groups, provide the best explanation for observed inequality.[53] The population of loan applicants will then fall into groups which do not correspond to their risk profile. Instead, membership in a group of loan applicants is determined by the lender's subjective preferences which do not (necessarily) correlate with credit default risk. For this reason, taste-based discrimination is not a rational reaction to the lender's situation of uncertainty. Rather, it can lead to inefficient rejections, if the lender's subjective preference for a specific set of loan applicants does not reflect their lower credit default risk.[54] In theory, taste-based discrimination is not expected to survive in a competitive market, given that it would indicate lenders are not objectively profit-maximizing.[55]

Empirical data is not always consistent with these theoretical forecasts. Bartlett et al. (2022) explore data on the Government Sponsored Enterprises (GSE), Fannie Mae and Freddie Mac. These GSE charge each loan a guaranty fee that depends only on the borrower's credit score and loan-to-value (LTV) ratio. In return, lenders are guaranteed against credit risk. The authors assume that interest rate differences between loans with identical credit score and LTV cannot reflect differential credit risk but must go back to some form of discrimination.[56] Using this strategy, they find a mark-up of 7.9 basis points for purchase mortgages and 3.6 basis points for refinance mortgages for Latinx and African American borrowers. For the lenders, this suggests than sometimes money is left on the table. For the borrowers, this results in these groups paying $765 million in extra interest per year.[57]

Fuster et al. (2022) find that "a large fraction of borrowers who belong to the majority group (…) experience lower estimated default propensities under the machine learning technology" but "these benefits do not accrue to some minority race and ethnic groups (…) to the same degree."[58] The authors show that, for minorities, better technology produces "winners" and "losers". Winners are

[51] Assumptions set by Fuster et al. (2022), p. 38.
[52] Arrow (1971); Phelps (1972); in the context of AI underwriting models: Hurlin et al. (2021), p. 6.
[53] Becker (1957); Becker (1993); newer models refer to stereotypes, assuming that loan examiners systematically underestimate the long-run profits of lending to minority applicants, Bordalo et al. (2016). Both models cannot explain why bias would persist in competitive markets, Dobbie et al. (2019), p. 7.
[54] See Bartlett et al. (2022), p. 32 for an example of unprofitable discrimination; further Hurlin et al. (2021), p. 6.
[55] Arrow (1971); Phelps (1972); Hurlin et al. (2021), p. 6.
[56] The authors have not explored if strategic pricing could be understood as a marketing tool by offering some customers very attractive prices to keep them interested.
[57] Bartlett et al. (2022), p. 31.
[58] Fuster et al. (2022), p. 8.

disproportionately white Hispanic and Asian. In Black and non-white Hispanic populations there are roughly equal fractions of winners and losers.[59]

Working with the services offered by Upstart, a US-based NGO ran a case-study. Applicants were construed as identical except for the college they had attended. Holding all other inputs constant, the authors of the study found that a hypothetical applicant who attended Howard University, a HBCU, would pay higher origination fees and higher interest rates over the life of their loans.[60] Similar results were observed for applicants who attended NMSU, a HSI.[61]

There is a variety of hypotheses to explain these empirical results. Some point to taste-based discrimination which they understand as persistent despite its economic inefficiency. Other explanations focus on strategic pricing as a reason for discrimination. Bartlett et al. (2022) speculate that their findings might have to do with minority borrowers being prone to less comparison shopping on average, having less experience or acting in a more urgent time frame.[62] Put differently: These authors explain discrimination not as a gain in subjective utility for the lender, due to taste-based preferences, but as a result of lenders targeting groups which - statistically - will be more vulnerable, hence open to predatory pricing. Explaining discriminatory results along those lines as statistical or taste-based discrimination or as strategic pricing has one immediate implication. If lenders compete in targeting vulnerable groups for either of these reasons, AI can help them to identify such groups.[63] I will come back to this much later when discussing how to move beyond anti-discrimination law.[64]

D. BIASES AND ALGORITHMIC UNFAIRNESS

AI underwriting models have raised high hopes if compared with either the limited list of input variables of traditional scoring bureaus or the biases and cognitive limitations of human credit officers.[65] At the same time, computer scientists caution against mistaking predictions based on big data and AI as "objective" forecasts.[66] Rather, the predictive power of an algorithm very much depends on its context. Using a spam filter as an illustration, Kim lists conditions for AI algorithms to be efficient: the target variable they look to is clear, the answer is a straightforward yes/no binary choice, there is a correct outcome, an error the algorithm makes can be used to refine it and unbiased

[59] Fuster et al. (2022), pp. 31–32.
[60] Student Borrower Protection Center (2020), methodology described at p. 16.
[61] Student Borrower Protection Center (2020), findings described at pp. 18-19.
[62] Bartlett et al. (2022), p. 32: "the fact that the relation between the rate differential and either credit score or realized default is minor suggests the income and LTV results may instead reflect something else, such as the correlation between income, financial sophistication, and a propensity to shop for rates"; similarly Gillis (2022), p. 39 ("personalized pricing"); Hurlin et al. (2021) "lack of fairness".
[63] Aggarwal (2021), p. 55 on lenders using AI to exploit borrowers' cognitive and behavioral biases.
[64] See below F.III.2.
[65] Sunstein (2019).
[66] For a critical discussion see Blattner/Nelson (2021); Kaminski (2019), p. 1538; O'Neil (2016); from the perspective of sociology: Burrell/Fourcade (2021), p. 224; Burk (2021), p. 1163: AI as a "prosthetic extension of human judgment"; Kiviat (2019a), p. 283; Kiviat (2019b), p. 1134; Kim (2022), p. 1 on the promise of an evidence-based approach.

data is readily available.[67] For credit scoring and underwriting, not all of these conditions are met. The target variable "credit-default risk" refers to non-performance on a loan and it is one possible target which the algorithm could optimize. It could look merely to a binary answer along the lines of: did the borrower perform on the contract or not? But the lender might have a different target variable in mind when maximizing his return on a payday loan and looking for maximum payback over a short period of time. There will often be shades of grey which are necessary to understand the risk a potential borrower poses. A predatory loan will be tougher to pay back than a standard, market-priced loan.[68] An unforeseen event, macro-economic, hence affecting everyone, or micro-economic, affecting only the borrower, might explain the non-performance.

Additionally, for credit underwriting, errors are often impossible to fix because there is no counterfactual data. The AI learns which loans it should not have offered because borrowers did worse than predicted (false positives). But it does not learn which loans would have been attractive, for the simple reason that they were not offered to the applicant.[69] Another explanation has to do with training data. Current AI underwriting algorithms use machine-learning and existing datasets on borrowers. Such datasets include their performance in repaying loans in the past. For these borrowers, the training data must also include alternative data of the type I have described above.[70] Based on the training data set, the AI "learns" to connect alternative data to performance. This allows the AI to judge and evaluate future applicants for a loan. The more similar the applicant's characteristics are to the characteristics of borrowers which were successful in the past, the better the score the AI attributes to this applicant: Yesterday's world shapes today's predictions.[71]

I. Yesterday's world and the credit-default risk target variable

It follows from there, that the quality of an AI's predictions is only as good as the match between the world according to the training data and the world as it is today.[72] If the training data reflects past inequality, then an applicant who shares features with a historically underserved group will be flagged as a higher credit risk than a comparable applicant who does not share the relevant feature (historic bias).

The fact that training data is, in this way, shaped by history has direct implications for how the AI builds its model.[73] Variables it finds for most candidates which were successful in the past will be accorded most weight, for instance a specific sex or race. Candidates whose profile does not include

[67] Kim (2022), p. 3, for a similar list see Kaminski (2019), p. 1539 (clear and mathematical objective, detailed and direct data, transparent inputs and code, easily verifiable outcome, fair and accurate output).
[68] See below D.I.; E.III.1.b.
[69] See below D.III; Kim (2022), p. 5.
[70] See above B.
[71] Mayson (2019), p. 2251: "look to the past as a guide to the future".
[72] Hellman (2020), p. 841 ("compounding injustice"), p. 842 (carrying forward injustice); Mayson (2019), p. 2251: "The premise of prediction is that, absent intervention, history will repeat itself".
[73] Blattner/Nelson (2021), p. 12 ("model bias").

the relevant positive variable will face a risk premium (majority bias).[74] In that way, AI underwriting models "lock us in" in yesterday's world.[75]

The same logic applies to variables which send a negative signal. The AI learns from historical data and singles out variables which have in the past signaled high credit-default risk. Applicants whose profile includes the risk-variable see their credit score sink. This happens even if a particular risk-variable does not reflect relevant details of the default situation across all applicants. The same is true if the observed risk-variable is less informative for some applicants if compared with others.[76] Take, for instance, an AI model which optimizes "credit-default risk". It predicts the probability of non-performance on loan and interest rate payments across borrowers. To illustrate the problem, assume that one group of the population is especially vulnerable to signing predatory loans. For some of these borrowers the reason for default might be the inability to pay the predatory interest rate. Hence, there is a good chance that the same person would have been able to perform, had she received the same loan under the same circumstances, but with a market-standard interest rate. However, if the model looks to non-performance only, a borrower who fails to pay back a loan with a predatory interest rate will be treated the same as a borrower who fails to pay back a loan with a market-standard interest rate.[77] If the group of borrowers which are (or were in the past) likely to accept predatory loans share a protected characteristic such as race, any applicant with that characteristic will be accorded a higher risk factor. A "colorblind" algorithm would be restricted from noticing the difference and – in that way – hurt, rather than help applicants of that race.[78]

II. Yesterday's world and transparency

The flaws due to algorithmic biases are not per se novel concerns or worries which are specific to AI underwriting. Traditional scoring models, for instance the US FICO score, with their limited number of input variables, provide a much cruder picture than today's AI underwriting models.[79] The FICO score methodology is shaped by path-dependent historical choices of relevant variables and of the balance struck between accuracy, search costs and market expansion.[80] It is common knowledge that applicants who do not fit the FICO profile find it hard to qualify for a loan.

---

[74] Barocas/Selbst (2016), p. 689; Gillis (2022); Graham (2021), p. 211; Langenbucher (2020); Burk (2021), p. 1163.
[75] See below D.III. on market forces contributing to ameliorating models.
[76] EDPB/EDPS (2021); Gillis (2022): "biased measurements"; Burk (2021), p. 1164 on the lack of context for late payments with the conclusion that credit scoring entail a "reproduction of social context", although it was "originally intended to help neutralize bias in lending".
[77] See for a similar problem in the context of predicting crime Mayson (2019), p. 2263 (three prior arrests in New Orleans were for a black man not unusual since "black men were arrested all the time for trivial things", by contrast, the same variable (three prior arrests) for a white man "was really bad news").
[78] See Hellman (2020), pp. 818, 848 et seq. on why the assumption that it is under US law prohibited for the algorithm to look into protected variables; Mayson (2019), p. 2259, making the point that "differential crime rates do not signify a difference across racial groups in "an individuals' innate" propensity to commit crime" and making an argument for allowing an algorithm to assess risk factors contingent on a protected characteristic; Langenbucher (2022), p. 364 on Art. 10 of the EU AI Act Proposal which allows the processing of data on protected variables if it is "strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to high-risk AI systems".
[79] In the EU, not all Member States have credit reporting and credit scoring agencies similar to the US. While Germany and the UK do, France does not and has lenders score applicants in-house.
[80] Bourrell/Fourcade (2021), p. 5.

However, from the perspective of the borrower who is subject to scoring, FICO-scoring will often be more transparent than AI and big data scoring. Traditionally, because of the limited number of variables, many reasons for a denial of credit will be obvious. A recent immigrant might face a risk premium for lack of a repayment history with US credit card companies. Filing a claim under the Fair and Accurate Credit Transactions Act (FACTA) will help him to learn about this concern by getting access to some of the factors underlying his score.[81] Understanding his score can guide him towards a solution, if he is in a position to influence the relevant factors.

This is not to deny the many flaws of traditional scoring systems nor to claim that changing the input variables which inform traditional scoring agencies is an avenue open to all groups of society.[82] The point made here is that AI underwriting offers less transparency than traditional scoring. Even if the ECOA and Regulation B require lenders to explain the reasons for denial of credit and to disclose the factors they used, it is unclear what this entails as to algorithmic scoring.[83] A lender might claim to meet the requirement by releasing computer code which is incomprehensible to most.[84] Even if some variables can be established, the model will often redundantly encode the same or very similar information in various variables.[85] Put differently: it might not help to bring in additional information, if the AI will extract the identical score by relying on other variables.

III. Yesterday's world and inaccurate data

These worries matter even more when combined with concerns about data quality.[86] Data can vary in its reliability across a population, for example if there is less data available for specific groups such as recent immigrants.[87] Additionally, the US CFPB has stressed that the use of alternative data, for instance stemming from social networks, increases the risk of inaccuracies. One of the reasons for that are quality standards. Social networks consist mostly of data uploaded by its users. Platforms do not engage in double-checking the accuracy of that data, much less scrutinize it along the lines of credit reporting agencies.[88] This concerns the applicant for a loan if the data used to evaluate him is inaccurate. It also impacts the AI model which learns from (partially) inaccurate training data. The more inaccuracies are hidden in big datasets, the more the AI "locks us in" in a world which does not even adequately reflect yesterday's world, much less today's.

---

[81] Not all variables are covered because scoring agencies and lenders are allowed to treat the scoring model, including the weight of each variable, as a trade secret. See Langenbucher (2020) comparing this to the (identical) German situation.
[82] This paper does not contribute to the extensive debate on the discriminatory potential of current scoring metrics, see: Bourrell/Fourcade (2021); Burk (2021), pp. 1163 et seq.; Citron/Pasquale (2014), pp. 11 et seq.
[83] 12 C.F.R. § 202.9.(b)(2).
[84] Information asymmetries between highly trained coders and consumers facing a denial could theoretically be overcome if a market for intermediaries develops, see Citron/Pasquale (2014), p. 29 for "creative customer relations" demystifying credit scoring via feedback and control mechanisms.
[85] "Flexibility" in the terminology of economists.
[86] See below F.I.
[87] See Mayson (2019) for the same problem leading to racial distortion.
[88] CFPB (2017), p. 17.

Market forces should eventually solve some of these issues. This is especially true for models with too many false positive results. The lender will at some point realize that his model does not adequately predict credit default risk and switch to a more powerful one. Of course, the road can be long and there is the risk of credit bubbles and concerns for financial stability to consider.

For models with too many false negative results, market forces will be less efficient in weeding them out. Some of this has to do with a specific feature of credit decisions, namely that there is no counterfactual data.[89] If the lender denies a loan to an applicant who is considered high risk, he will never know whether the better decision would have been to grant that loan. Accordingly, the AI will never learn which borrowers it denied a loan although it should have offered them one.[90] This is different in other use cases of AI. Imagine a doctor, using AI to scan melanoma. The AI might wrongfully overlook a critical result in year one. If the same patient returns in year two and three, the AI might eventually put his melanoma in the correct category. In this way, it expands its data and learns that it should have categorized the melanoma differently in year one. For credit decisions, this counterfactual data is not available, hence, the AI cannot learn from the false negative decision.

E. SCOPE AND LIMITS OF ANTI-DISCRIMINATION LAW TO COPE WITH AI SCORING AND CREDIT UNDERWRITING

Understanding algorithmic discrimination in credit scoring and underwriting does not start with a clean slate. Relevant court cases concern areas as diverse as employment, housing, bail, and credit scoring.[91] Legal rules and principles address the employer's decision on whom to hire, the landlord's criteria for selecting tenants, the judge making bail decisions and the lender denying credit. While there are common core principles of anti-discrimination law, each of these scenarios has in the past called for a different balance between competing interests and values. This is perhaps most pronounced for credit scoring, where distinguishing different sets of loan applicants based on sophisticated statistical risk models has been an institutionalized part of making a credit decision. Additionally, macro concerns of financial stability and protecting borrowers from over-indebtedness have for decades fueled the search for powerful predictors of credit default risk.

The US has in the late 1960s started to regulate fair lending. The EU has only in its EU Consumer Credit Directive/2021 explicitly addressed discrimination in lending.[92] So far, there have been EU rules prohibiting discrimination based on race, ethnic origin and sex, but they have required the loan to be a product or service available to the general public irrespective of the borrower's personal

[89] See above D before I.
[90] Kim (2022), p. 5. A problem which raises similar concerns has been described for predicting future crime, see Mayson (2019), p. 2252. "criminal justice risk-assessment tools purport to predict future crime. But that is not actually what they predict. They generally predict future arrest".
[91] For an overview see Kissinger et al. (2021).
[92] See above A.

situation. Beyond that, fair lending has remained a question of Member State law.[93] The EU Consumer Credit Directive/2021 introduces an explicit anti-discriminatory rule.[94] It protects consumers legally resident in the EU from discrimination on grounds of nationality, place of residence, sex, race, color, ethnic or social origin, genetic features, language, religion, belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation.[95]

Anti-discrimination laws do not distinguish between taste-based and statistical discrimination, nor do they allow for statistical discrimination on grounds of a protected attribute.[96] Some of the strategies an economist views as efficient when faced with uncertainty will be impermissible under the law. Put differently: An efficient underwriting decision can be unlawful even if it reflects an unequal distribution in the world.

Anti-discrimination law is triggered by input to a decision-making process. It starts from the assumption that such input comes in the form of different attributes or practices.[97] I refer to them as "building blocks" of a decision. There are outright prohibited building blocks, and facially neutral ones. Impermissible building blocks must not be used, even if they are of direct empirical relevance.[98] At first glance, facially neutral ones can be used. However, in some situations even facially neutral attributes might still be "suspicious", as it were. This is the case if they are "fair in form, but discriminatory in operation".[99] A well-known example is provided by the US Supreme Court decision in *Griggs,* where the score in an intelligence test was decisive for the position as a manual laborer. While this did not specifically look to age, sex or race, it triggered a discriminatory result. The same was true for the facts of *Smith,*[100] where years of experience in a job lead to a proportionately lower pay raise. Both, the score in an IQ test and job experience are facially neutral variables. However, they can operate in a way which masks the true reasons underlying the decision. Such is the case if the seemingly neutral attribute is picked *because* it correlates, for instance, with age or with race.[101]

This paper puts a spotlight on the role of such building blocks. Traditional doctrine requires these building blocks to form a chain of causation which eventually leads to a decision. The law's role is to carefully examine each building block and to determine whether the decision would have looked different, if one unlawful building block was removed from the chain of causation: Would the person

---

[93] See Langenbucher (2020). In Germany, §§ 19, 20 *Allgemeines Gleichbehandlungsgesetz* (General Act on Equal Treatment) have transposed EU Directives.
[94] Proposal for a Directive on Consumer Credits of 30 June 2021, COM(2021) 347 final.
[95] Art. 6 of the proposal, see already recital (45) of the current Consumer Credit Directive 2008/48/EC.
[96] See above C.
[97] On input see Berman/Krishnamurthi (2021), p. 99; Koppelman (2022). p. 10 (the latter criticizing the former, but in agreement about this basic point).
[98] See Gillis (2022), pp. 49, 51 for the claim that this blurs the distinction between anti-discrimination law and affirmative action.
[99] Griggs v. Duke Power Co., 401 US 424 (1971) at p. 431.
[100] Smith v. City of Jackson, 544 U.S. 228 (2005).
[101] This is not to understand the law as blind to economic explanations of discrimination as a statistical phenomenon. Some of the reasons leading to statistical discrimination can show at a later stage of the reasoning process, namely where disparate impact doctrine allows for a legitimate business defense. See below E.III. and see Gillis (2022), p. 48.

have been hired if the employer had not known her sex? Would the landlord have signed the lease if he had been ignorant as to the tenant's race? But also: Would the same employees have been hired if the IQ test was not run? Would the same bonus payments have been made, if years of job experience were not considered?

With improving technology this core notion of anti-discrimination law faces a novel challenge. Economists such as Fuster et al. (2022) predict that we will soon reach a situation which makes prohibiting the use of specific variables ineffective.[102] Big data furnishes a universe of different variables. Machine learning algorithms unearth innumerable correlations between those variables. Depending on the type of machine learning model employed, its human user might often not be able to identify salient variables or core correlations. Instead, we are looking at an opaque bundle of building blocks which drive a decision. As I will argue below, it is difficult to square this with the conceptual underpinning of anti-discrimination law, based on distinct building blocks, which form a neat chain of causation leading up to a final decision.

I. Three hypothetical lenders

One set of building blocks, for instance sex, race, or religion, are considered directly discriminatory. Anti-discrimination laws react by prohibiting decisions "because of" [103] or "on grounds of"[104] a protected attribute. However, reasons, intentions and motives are not always a straightforward phenomenon. Attributes which lead to the denial of a loan may, for instance, reflect a discriminatory taste-based preference and at the same time provide a useful signal for statistical discrimination.[105] This is especially likely if many members of a population have shared the same taste-based preference, if this has triggered credit rejections in the past and if this has today led to clusters of groups with high credit default risk.[106] Such situations might explain why lenders claim they do not discriminate *because of* any bias on their side but *because of* the efficiency arguments underlying statistical discrimination.[107] As we will see in more detail below, an argument along those lines will not usually exclude liability of the lender. Disparate treatment laws do not require a protected attribute to be the sole building block. Rather, decisions where only one out of a variety of variables is impermissible are considered disparate treatment, if they are a necessary building block along the chain of causation. [108]

Building blocks which are not outright prohibited might still be "suspicious", even if they are facially neutral. Dealing with these is where disparate impact doctrine comes into play. Disparate impact doctrine deals with decisions which are motivated by a facially neutral attribute, but which still trigger

[102] Fuster et al. (2022), p. 8; along similar lines: Gillis (2022).
[103] See the ECOA and the FHA for the lending context.
[104] Art. 2 (a), (b), 4 para. 1 (a), (b) EU Directive 2004/113/EC.
[105] CFPB (2017), p. 19; Dobbie et al. (2019) p. 1; Arrow (1971); Phelps (1972).
[106] Sunstein (2019), p. 509: algorithms using factors which are "an outgrowth of discrimination"; Gillis (2022), p. 18: "biased world".
[107] See Sacksofsky (2017) on German law's understanding of "because of".
[108] See below II.1.

a disproportionately disadvantageous outcome for protected minorities. Using this attribute is impermissible unless the decision-maker can establish justificatory reasons such as a business defense.

To illustrate, I introduce three hypothetical lenders, claiming that their AI underwriting model does not violate anti-discrimination laws. For the purposes of this paper, I assume that the lender's optimization goal[109] is to assess credit default risk.[110] I assume further that the lender is the one developing the model. There are additional scenarios in practice, for instance an AI credit scoring agency furnishing a report to the lender or a Fintech platform screening lenders but partnering with a bank to originate the loan. Such scenarios raise questions of liability of each involved party which are beyond the scope of this paper.

The first hypothetical lender reasons as follows: "Yes, I have trained the AI model to include sex in the observable variables I use to calculate credit risk. However, this is just one of the many observable variables I use. I include it because, statistically, sex is a good indicator for credit default risk."

The second lender claims: "I understand denying credit because of sex is impermissible. Therefore, I use an input-control procedure which makes sure that no protected characteristic enters my AI model."[111]

The third lender submits: "I don't even look at variables and, frankly, don't care much. My underwriting model is based on a huge number of variables. Also, I use a black-box model and just go along with whatever it suggests. So far, this has made good business sense".

Arguably, these hypothetical lenders raise different but related concerns. The first lender's claim has to do with mixed motivational elements and causation if we understand him as saying: "sex was not the sole cause of my denial of credit, it was just one of the variables I use".

The second lender's case is more complicated. Throughout the paper, it has become clear that the combination of big data and AI will lead to proxies standing in for protected attributes. Do proxies trigger disparate treatment liability, at least if they correlate narrowly with a protected attribute? Is it relevant whether the lender intentionally chose these proxies? Or is disparate treatment inextricably tied to a protected attribute and will the lack of a protected variable trigger (only) disparate impact liability?

The third lender is what Fuster et al. (2022) might have in mind when they claim that with technology improving it will become ineffective to prohibit the use of certain variables.[112] Arguably, decisions produced by a vast array of variables, not even necessarily known to the human who employs the AI model, are hard to square with received anti-discrimination law. The reason for this is that one of the

[109] On these see below F.III.2.
[110] Optimization goals such as targeting vulnerable groups for strategic pricing are discussed further below at F.III.2.
[111] See for this strategy in Fintech lending: Di Maggio et al. (2021), p. 4; for the case of Upstart: Langenbucher/Corcoran (2022), p. 143.
[112] Fuster et al. (2022), p. 8.

cornerstones of anti-discrimination doctrine is the concept of distinct building blocks forming a chain of causation. If these become hard to pin down and easily interchangeable without altering the result, this cornerstone loses significance.

II. Limits of Disparate Treatment

In the US, discrimination in a lending context is addressed by the ECOA and the FHA. In the EU, a common rule on discriminatory lending is envisaged in the EU Consumer Credit Directive/2021.[113] Against this background, this section focuses on US law to illustrate the logic behind anti-discrimination law.[114] Both, the ECOA and the FHA prohibit decisions which are motivated by a protected characteristic, the FHA in the context of mortgages, the ECOA for more general access to credit. While the conceptual framework is straightforward, proving a discriminatory building block is often difficult.

1. Establishing a Disparate Treatment Case

There are generally two regimes available to make a disparate treatment case, both developed in Title VII which covers employment discrimination. The first regime, following *McDonnell Douglas*, is focused on strategies which allow to prove the existence of a discriminatory motive.[115] The plaintiff might be able to establish overt or other direct evidence.[116] However, given the awareness of many employers (or lenders) of anti-discrimination laws, such evidence might often be hard to come by. This is even more likely if subconscious motives played a role in decision-making. Against that background, two further strategies allow for inferential proof. Individual inferential proof is common in Title VII cases. A plaintiff will need to establish a prima facie case by showing that she is a member of a protected group, was qualified for a position, was rejected and the position remained open. If she succeeds, the defendant must establish a legitimate non-discriminatory explanation. To do so, he must show that there was no discriminatory motive at play.[117] He does not have to prove that the reason he advances is true. Instead, it is only a burden of production. The plaintiff may react by attempting to prove that these reasons are pretextual. Group or systemic inferential proof is another way to make a disparate treatment case. Plaintiffs use statistics to prove a pattern and practice which reveals that their group is underrepresented.[118] Defendants may bring in different statistics or put forward a legitimate nondiscriminatory explanation to rebut.[119]

---

[113] See above A.
[114] Arguably, the logic of an anti-discrimination case in a credit context will under EU law follow rules similar to established cases, e.g. in an employment context. The argument I make on causation is general enough to apply to both, EU and US law.
[115] McDonnell Douglas Corp. v. Green, 411 U.S. 792 (1973).
[116] Equivalent in Germany: Sacksofsky (2017), p. 73.
[117] Similar in Germany: Sacksofsky (2017), p. 73.
[118] Similar in Germany: Sacksofsky (2017), p. 84 on § 22 *Allgemeines Gleichbehandlungsgesetz* (General Act on Equal Treatment).
[119] See below E.III. for the role of statistics in making a disparate impact case.

The second regime to make a disparate treatment case concerns a mix of factors. It has been applied in situations where it is not in doubt that a discriminatory element contributed to the decision, but a defendant still disputes causation. In line with a disparate treatment claim rooted in causal proof, courts will remove each building block and check whether the decision would have come out differently. An example is provided by the US Supreme Court decision in *Price Waterhouse v. Hopkins.* The defendant's decision to let a woman go was at stake.[120] She claimed she was fired because of her sex. The defendant had to establish that its legitimate reasons standing alone would have led to the same decision.

In *Manhart*, the Supreme Court held that an employer's policy of requiring women to make larger pension fund contributions than men violated Title VII. There was no doubt about an unlawful factor since the policy specifically targeted women. Still, the employer argued that he had no discriminatory intent and did not treat women differently *because of* their sex. Rather, actuarial logic dictated a "life-expectancy adjustment".[121] It is a claim any economist would have embraced, pointing to the logic of statistical discrimination. The US Supreme Court did not follow this reasoning. Instead, it was sufficient to establish that one impermissible attribute was a building block towards the decision. Removing the attribute "female" from the set of variables, so the Court held, would have led to a different, non-discriminatory outcome.[122] In a comparable case, the European Court of Justice rejected the claim of insurance companies which had argued statistics and actuarial logic required an adjustment of fees for women.[123]

2. The first Hypothetical Lender: No Statistical Efficiency Defense in Disparate Treatment Cases

The first hypothetical lender I described above explicitly used an impermissible attribute. He made two claims to justify its use. First, he suggested that it is but one of the many variables he feeds into his model. Second, he stressed that he had no discriminatory intent but just followed business logic.

a) Mixed motives

The first claim has to do with causation. The lender would be right if disparate treatment required the protected attribute to be the sole step in the chain of causal elements towards the decision. In *Price Waterhouse v. Hopkins*, the Supreme Court rejected this argument, given that the text of the statute did not read "solely because of".[124] For Title VII cases, it is established practice that plaintiffs must

[120] Price Waterhouse v. Hopkins, 490 U.S. 228 (1989).
[121] City of Los Angeles v. Manhart, 435 U.S. 702 (1978); the ECJ followed the same logic: ECJ ECLI:EU:C:2008:397; Sacksofsky (2017), p. 73.
[122] The argument differs from the Pricewaterhouse case. In that decision, the sex of the woman was one factor. However, the employer introduced a hypothetical line of causation. He argued that he could have reached the same decision with a different, non-discriminatory motive in mind, see Koppelman (2022), p. 14. In Manhart, the employer argued that his "real" motive was non-discriminatory but simply a short observable variable (sex) to predict an unobservable variable (life-expectancy).
[123] ECJ ECLI:EU:C:2011:100.
[124] Price Waterhouse v. Hopkins, 490 U.S. 228 (1989).

prove a form of causation[125] which looks to one out of various building blocks of the decision.[126] Events can have multiple "but-for causes" and the plaintiff's case is successful if he can show that the protected attribute was one of the motivating factors. Under this standard, plaintiffs have to show that removing the relevant building block changes the outcome.[127]

If the ECOA and the FHA adopt the standard to prove causation in the way Title VII does, a lender could not escape liability by citing a variety of observable variables which he included in his decision. So long as a protected attribute was one "but-for cause", it is enough to trigger the prohibition. Arguably, the wording of the ECOA and the FHA support this line of reasoning. Section 2000e-2(a)(1) of Title VII stipulates that it is unlawful to discriminate "because of" a protected attribute. It is this term which the *Bostock* Court has invoked to apply what it understands as the but-for standard of causation. Similarly, the FHA speaks of discrimination "because of" protected characteristics and the ECOA makes it unlawful to discriminate "on the basis of" certain protected attributes.[128] None of these statutes require that the outcome was reached "solely" because of the protected attribute.[129] This suggests that both statutes can be read along the same lines as Title VII.

Today, Section 2000e-2(m) of Title VII explicitly allows for a complaining party to demonstrate that the protected attribute was "a motivating factor for any employment practice, even though other factors also motivated the practice". The statute's text was changed to its current wording after the decision in *Price Waterhouse v. Hopkins*.[130] Congress did not change the wording of the ECOA and the FHA. Arguably, this cannot be construed as ruling out a mixed-motive test along the lines of Title VII's mixed-motives test. However, Title VII'S Section 2000e-2(b) includes a follow-up rule. It addresses a situation where the defendant can establish that he would have reached the same decision in the absence of the motivating factor. The rule still allows for declaratory relief, injunctive relief and attorney's fees and costs but it limits the relief available (disallowing damages, reinstatement and more). The ECOA and the FHA lack a provision along those lines, hence, it remains an open question whether the limited relief is allowed in ECOA and FHA cases.

For the first hypothetical lender, we learn that he does not escape liability by claiming that sex was not the sole cause of his decision. However, a plaintiff would still have to establish causation. When AI models are used this entails proof that the lender would have reached a different outcome if there had been an input restriction on the variable sex. For most algorithms, this will be hard to show because many variables correlate with sex. Even if the lender restricts input, due to redundant encoding, the outcome will often remain the same.

---

[125] See Dembroff/Kohler-Hausmann (2022), pp. 74 et seq. for a critique of applying causation standards borrowed from torts to anti-discrimination law.
[126] Berman/Krishnamurthi (2021), pp. 100 et seq.; Eidelson (2022), pp. 797 et seq. on reading the term "because of" as "by reason of", rather than looking to anything which contributed to the outcome in any way.
[127] Bostock v. Clayton County, 590 U.S. (2020), p. 6; Dembroff/Kohler-Hausmann (2022), p. 58.
[128] EU law's close analogue reads: "on grounds of".
[129] 15 USC Chapter 41 § 1691; 42 USC § 3604.
[130] See Berman/Krishnamurthi (2021), pp. 99 et seq. on the decision and on Congress following up by amending Title VII to encompass a mix of motivating factors.

b) Intent

The second claim of the hypothetical lender stressed that the protected attribute correlates with high credit default risk and that this is the only reason why he includes sex. Framed in this way, the first hypothetical lender is not necessarily talking about causation.[131] Rather, he has discriminatory motives or intent in mind. He claims that his decision should not be understood as taken *because of* an impermissible motive in the way a taste-based discriminator proceeds.[132] Instead, he screens potential borrowers for sex, race, or similar attributes *because of* an entirely different reason, for instance with statistical discrimination in mind. This different reason could then be the useful contribution these attributes made for reducing the lender's uncertainty about other (unobservable) attributes. Put differently, he submits that he had no discriminatory intent whatsoever, but just implemented his business strategy.

Disparate treatment doctrine is not open to this line of argument. If the impermissible attribute (not a proxy) was one building block of the decision, and the decider was aware of that, the law will understand the lender as having discriminated *because of* the protected attribute. As illustrated by *Manhart*, proving discriminatory intent is not a necessary requirement for making a disparate treatment case. Put differently: If the protected characteristic is one building block of the causal chain, intent to discriminate in the form of taste-based discrimination is not required. I will get back to this point when discussing proxies.[133]

By the same token, there is no such thing as a "statistical-efficiency defense" available in disparate treatment cases.[134] Although actuarial statistics in *Manhart* suggested a risk premium for women as a group, this did not help the defendant. In *Bostock* the Court confirmed its earlier reasoning. It referenced *Manhart* and the inadmissibility of a "life-expectancy adjustment".[135] This suggests that it would be just as futile for a lender to call a policy which discriminates against women a "credit-default risk adjustment", even if looking to sex allowed for efficient statistical discrimination.

Arguably, the Court would reach the same conclusion for a disparate treatment case under the ECOA or the FHA. Just like the employer in *Manhart*, a lender might wish to bring in statistics, showing why a risk premium should attach to, for instance, sex, race, or age. He might wish to add that this risk premium motivated his decision, which he understands as non-discriminatory. If the Court decided along the reasoning in *Manhart*, it would not be hearing this argument. As soon as an

[131] By contrast, Dembroff/Kohler-Hausmann (2022), pp. 74 et seq. link the question to causation when they claim that causation in anti-discrimination law requires a "Normative Showing" along the following lines: "If not for the defendant's *discriminatory* conduct, policy, motive, or intent, would the plaintiff have experienced this employment practice or loss?"
[132] See above C for taste-based and statistical discrimination.
[133] See below at E.II.3.b.
[134] See for a comparison to disparate impact's business defense: Barocas/Selbst (2016), p. 713 referencing 42 U.S.C. § 2000e-2(k)(2) (2012).
[135] Bostock v. Clayton County, 590 U.S. (2020), p. 2.

unlawful characteristic appears in the chain of building blocks leading to the decision, a disparate treatment case is established, irrespective of any additional discriminatory intent.

As a doctrinal point about disparate treatment doctrine, it is worth mentioning that legislation such as the ECOA and the FHA send a clear message about a trade-off between business efficiency and equal treatment: There will be no such trade-off.[136]. In practice, lenders will often use a facially neutral variable which correlates with a protected characteristic: a "proxy". This brings me to the second hypothetical lender.

3. The Second Hypothetical Lender: Proxies

The defense put forward by the second hypothetical lender is more intricate than the one of the first hypothetical lender. He claims to escape liability because he controls input to his model, making sure no protected characteristic enters. A court, so this second hypothetical lender submits, which examines the chain of building blocks will not find one impermissible attribute among these elements.

How to handle proxies has been a classic concern of anti-discrimination doctrine. This has to do with one of the foundational assumptions of anti-discrimination law this paper wishes to highlight.[137] Above, I have described the concept of building blocks leading to a decision as one of the core tenets of anti-discrimination law. Some of these building blocks are outright unlawful, some are facially neutral but suspicious, some are altogether harmless. With a human decision-maker in mind, most courts and scholars follow the ground rule that proxies do not usually qualify as an impermissible characteristic. At the same time, some situations are understood to require exceptions to this general rule. This is the case, for example, if the proxy is identical to the protected characteristic and differs in name only or if the proxy serves as a pretext for the real motive.[138] Implicit in this approach is the understanding that there is a limited number of proxies, that a human decision-maker has intentionally picked the relevant proxy and that his motives for doing so will guide the way when evaluating his decision.[139]

With the advent of big data and AI models, these implicit assumptions do not necessarily hold. The number of proxies will grow immensely. This is true for individual proxies which an AI model finds to correlate with a protected attribute. It is even more evident for bundles of variables an AI identifies in the sea of big data which, when combined, allow to predict the probability of a person sharing a protected attribute. To illustrate, consider marketing researchers who have found numerous strategies to predict age or sex, based on online behavior, mobile phone services, natural language processing or twitter usage.[140] Often, data analytics of this type will work with a bundle of variables, finding

[136] Sunstein (2019), p. 504 applauds algorithms to reveal the extent of this trade-off.
[137] See above E.
[138] See below E.III.2.
[139] Kim (2022), p. 15.
[140] Illustrated by Al-Zuabi et al. (2019).

patterns and correlations to predict sex or any other protected characteristic with a high or very high probability. Discrimination by proxy will develop from an exception to the standard case.

For now,[141] let us assume we are looking at a single proxy which is a good predictor for a protected characteristic. If such a variable is used by a lender's AI model, does it qualify as a protected characteristic? Does it matter how accurate the model's forecast is?

a) The taste-based discriminatory lender

A straightforward case is presented by a lender who "knowingly and purposefully bias(es) the collection of data" to satisfy his discriminatory taste-based motivation.[142] Barocas and Selbst have called strategies along those lines "masking". If the plaintiff succeeds in proving "masking" behavior, the defendant violates anti-discrimination laws. "The entire idea of masking", they explain, "is pretextual".[143] The argument has intuitive appeal, arguably because the lender's behavior rings of circumventing legal rules. If the true motive for a decision is the protected attribute, the defendant should not be able to escape liability by hiding behind a pretext.

Masking is not a new phenomenon, but algorithms using big data bring about new ways to conceal true intentions. From a doctrinal point of view, masking case are easy cases. The challenge lies not with applying disparate treatment doctrine, but with establishing proof of masking behavior. Even for inferential systemic proof,[144] courts might in the future face a battle of competing AI models to bring in statistics. Plaintiffs will face the task of showing that the discriminator deliberately chose a variable to mask his discriminatory intent.

b) When does masking end and disparate impact start?

The line between concealing discriminatory intent and unsuspectingly using a facially neutral characteristic (which entails only disparate impact liability) can be a fine one. What about a lender who understands that her AI model produces distinct sets of applicants and that group membership in these sets tracks attributes of protected communities? Does awareness alone make it a disparate treatment case? Does it depend on how narrowly the facially neutral variable correlates with a protected characteristic?

Barocas and Selbst (in an employment context) suggest that "intent is clear, if the employer continues *because* he liked the discrimination produced".[145] They extend liability to situations where employers "preserve the known effects of prejudice in prior decision making".[146] This is in line with the

[141] See below E.III. for bundles of variables.
[142] Adams-Prassl et al. (2022), p. 4; Kim (2022), p. 15.
[143] Barocas/Selbst (2016), p. 699.
[144] See above E.II.1.
[145] Barocas/Selbst (2016), p. 699.
[146] Barocas/Selbst (2016), p. 692.

traditional understanding of proxy discrimination as one form of intentional disparate treatment.[147] It is also very similar to the taste-based discriminatory lender's masking. Redlining is a paradigm example in the US loan context:[148] In certain areas, ZIP-code is a proxy for race. A lender who deliberately uses ZIP-codes as a building block towards his decision does not escape disparate treatment liability.[149] At the core of understanding these situations as cases of disparate treatment lies the motivation and intent of the discriminator: he "likes" the effects he produces, or he keeps well-known consequences of his model.[150]

At the same time, Barocas and Selbst submit that "deciding to follow through on a test with discriminatory effect does not suddenly render it disparate *treatment*".[151] This seems to suggest that awareness of a correlation between the facially neutral variable and the protected attribute alone is not enough to trigger disparate treatment liability. As Hellman explains: "The Supreme Court has insisted that a screening tool must have been adopted "because of" the disparate impact and not merely "in spite of" these foreseeable consequences in order to give rise to strict scrutiny".[152] It is important to note that the extent to which an individual proxy can stand in for a protected attribute is a question of degree. A proxy might correlate very narrowly with a protected characteristic – as was the case for redlined regions and race. For other proxies, the correlation will be less significant.

Again, drawing the line between "because of" and "in spite of" is not a novel concern, however, the use of sophisticated AI models compounds existing difficulties. Before, when lenders picked their building blocks for an underwriting decision, they deliberately chose the ones they considered relevant. Plaintiffs had to establish discriminatory intent in the context of that decision. With algorithmic decision-making, it is not necessarily the human lender who chooses the variable. This changes what plaintiffs must scrutinize when looking to prove intent. Instead of investigating the decision to pick the proxy, plaintiffs must explore the state of mind of the lender who realizes the discriminatory potential of his model. A disparate treatment case is established if the plaintiff can show that the lender kept using the algorithm because of the discriminatory potential. The awareness of a correlation alone, even if it is a narrow one, does not make the case one of disparate treatment.

---

[147] Hellman (2020), p. 851: "if a facially neutral classification (i.e. not race, sex, or some other protected trait) is used deliberately as a proxy for a protected characteristic, the use of the so-called "facially neutral" (or non-protected) classification also gives rise to heightened judicial review". Id., pp. 856 et seq. on racial classification, explaining why not all racial classifications are subject to strict scrutiny

[148] Under EU law there is little discussion on redlining. See for an understanding along indirect discrimination Dzida/Groh NJW 2018, 1917.

[149] Prince/Schwarcz (2020), p. 1257; Campbell/Smith (2022), p. 9 present redlining first as an example for indirect discrimination, however, see p. 14 where the authors claim it is direct discrimination because the protected characteristic was considered in deciding. Adams-Prassl et al. (2022), p. 12 use a UK case which presents a similar problem concerning age discrimination.

[150] Authors seem to disagree about requiring intent. Arguably, some disagreements have to do with imprecisions on what is meant by "intent": (i) discriminatory goals along the lines of taste-based discrimination described above or (ii) deliberately looking to a protected attribute as one motivational building block, but for other reasons, for instance statistical discrimination (as was the case in Manhart). See Mayson (2019), p. 2240 for requiring type (ii) intent. For UK authors reading US, but not EU and UK law as requiring intent type (i): see Adams-Prassl et al. (2022), p. 6; Campbell/Smith (2022), p. 3.

[151] Barocas/Selbst (2016), p. 699.

[152] Hellman (2020), p 852.

The Fintech lender Upstart provides a good illustration for a lender who, arguably, was aware of a troubling correlation and still wished to escape rather than cement discrimination.[153] Upstart's underwriting model uses various variables. Importantly, it only processes variables in concert, not in isolation.[154] Educational background is one of the variables used and Upstart claims that this allows to grant more loans to minority groups than lenders working only with traditional FICO scores. This inspired the mystery shopping exercise I described above.[155] The authors found that under Upstart's model borrowers with an educational background in a historically Black or a Hispanic/Latinx institution paid a significant penalty on both, origination fee and interest rates, if compared with borrowers who attended NYU. While Upstart did not use race as a variable, it would be surprising for a US lender to be unaware of a correlation between race and education.[156] Still, there was no indication of discriminatory intent on the side of Upstart. Rather, including education (among other variables) contributed to a more granular prediction of credit default risk, to finding more "invisible primes" and to (in the aggregate) originating more loans.[157]

Against this background, we see a first answer emerging for situations where the lender is aware of a correlation between his motive and a protected characteristic but is not intentionally concealing his true discriminatory intent. If the protected characteristic itself is a building block of the chain of causation, disparate treatment liability ensues. Discriminatory intent is not required.[158] By contrast, if the lender uses a proxy (a facially neutral variable which correlates with a protected characteristic), disparate treatment liability requires him to intentionally hide behind the seemingly facially neutral

[153] Langenbucher/Corcoran (2022), p. 152.
[154] CFPB (2017), p. 4.
[155] See above C.
[156] This assumes that Upstart was not hiding its true motivational element of discriminating on the grounds of race. If this were the case, we would be faced with a masking situation as explained in the preceding paragraph.
[157] CFPB (2017), p. 6.
[158] Adams-Prassl et al. (2022), p. 6 claim that unintentional discrimination can be direct under European law but not under US law. However, cases such as *Manhart* show that an intention to discriminate is not necessary if the protected attribute is one building block of the decision. For aligning US and EU law as proposed here: Hacker (2018), p. 55; Wachter et al. (2021), p. 41; Zuiderveen Borgesius (2020), p. 31.

variable. There are some exceptions to this rule for discrimination by proxy such as redlining.[159] In practice, this will often mean that only a disparate impact case is available.[160]

c) One variable "necessarily entails" a protected characteristic

Somewhat fuzzy situations arise if a lender uses a proxy which not only correlates with a protected attribute but is understood to be somehow implied by it.[161] Pregnancy is a classic example.[162] Neither US nor European law had explicitly listed the term "pregnancy" as a protected attribute. A textualist reading would expect courts to address the issue as one of proxies (facially neutral variables which correlate with a protected characteristic). This is indeed what US courts did in the 1970s. In *Gilbert*, the US Supreme Court held that exclusion of pregnancy from a disability benefits payments plan was not based on "sex".[163] Congress had to amend Title VII to extend its protection to pregnancy, closing an apparent gap. The European Court of Justice, following more purposive principles of interpretation, found that pregnancy is "inextricably linked" to the female sex. A refusal to employ an applicant due to pregnancy, so the Court reasoned, can only concern women.[164] Rather than have plaintiffs wait for a legislative amendment, the Court proceeded with a broad reading of the term sex. Pregnancy was addressed as an attribute "inextricably linked" to sex.

The US decision in *Bostock* arrived at a similar result when applying Title VII's prohibition of discrimination on grounds of sex to discrimination because of sexual orientation. Claiming that this followed from a textualist interpretation of the term sex,[165] the Court argued that it is impossible to

[159] There is no accepted standard for deciding which case qualifies as a discrimination by proxy situation and much of this, arguably, is cultural. In the UK, the term "pensionable age", so a UK Court in James v Eastleigh Borough Council (1990) 2 AC 751 held, had become a shorthand expression which refers to the age of 60 in a woman and to the age of 65 in a man. What ZIP codes represent in the US as to race, "pensionable age" represents in the UK as to age. In my view, this does not support the broader view of Adams-Prassl et al. (2022) that EU law's scope of direct discrimination is broader than US law's disparate treatment. Additionally, exceptions to this rule are sometimes considered if a proxy correlates 100 % with the protected attribute. The overwhelming majority of courts and scholars are not open to crossing the 100 % line. See the UK case of Lee v. Ashers Baking Company Ltd and others [2018] UKSC 49 ("exact correspondence test"); Sacksofsky (2017); for different reasons: Campbell/Smith (2022), pp. 10 et seq. By contrast relaxing the 100 % benchmark: Adams-Prassl et al. (2022) (but without a clear new benchmark). Arguably, exceptions made for individual proxies which correlate 100 % with a protected attribute are not the most relevant case for AI underwriting models which, more often, rely on a bundle of variables. Additionally, while predicting the existence of a protected attribute with a 100 % probability can be a useful goal for use cases such as targeted advertising, this is not a necessary goal of an underwriting model (unless we are faced with the devious discriminatory lender investigated in the context of masking, see above at E.II.3a).

[160] For EU law see Hacker (2018) p. 55; Wachter et al. (2021), p. 41; Zuiderveen Borgesius (2020), p. 31; for a broader understanding under EU law see Adams-Prassl et al. (2022).

[161] See Adams-Prassl et al. (2022), p. 12: "inherently discriminatory"; Krishnamurti/Salib (2020): "Conceptual Causation"; referring to the latter: Berman/Krishnamurthi (2021), pp. 88 et seq.; discussing "being a mother" as a "true subset of one sex" on p. 105.

[162] For the sake of this example, I do not address the situation of transitioning persons where a man might become pregnant, see Sacksofsky (2017).

[163] General Electric Co. v. Gilbert, 429 U.S.125 (1976), p. 149.

[164] ECJ ECLI:EU:C:1990:383 (note 2).

[165] The debate whether the Court's result in Bostock can indeed be explained under a textualist approach is outside the scope of this paper; see Berman/Krishnamurthi (2021).

discriminate against homosexual or transgender persons without first ascertaining their sex. Decisions based on homosexuality or transgender identity were held to "necessarily entail" sex.[166]

So far, few court cases speak about proxies which "necessarily entail" a protected characteristic and the conceptual fuzziness these cases introduce to the separation between prohibited attributes and proxies has not yet caused too much concern.[167] With big data and AI models this could change. It is quite probable that we will face many novel proxies: variables which allow to predict protected characteristics with a very high probability. Marketing research and targeted advertising firms have long worked on this task. They have found individual variables, such as first name, height or google search history, which allow to forecast sex with some probability.[168] Predictive power is stronger the more (bundles of) variables we use and the more correlations the algorithm finds.

This has implications for the core of anti-discrimination law, triggered by distinct motivational building blocks. The fuzziness which "necessarily entailing" variables brought about is multiplied: If a specific bundle of variables allows with a very high probability to predict a protected characteristic, can we say that the bundle in its entirety "necessarily entails" a protected attribute? Does it make sense to decide this novel question along the lines of narrow precedents on pregnancy and sex identity? This brings me to the case of the third hypothetical lender who is neither interested in understanding individual proxies, nor in bundles of proxies or their relationship to protected attributes. He uses his AI model as a black box, helpful for predicting credit default risk.

4. The Third Hypothetical Lender

The third hypothetical lender will arguably be both, the most frequent and the most troubling case. His motivation is business profit. He does not mask any discriminatory intent, nor does he qualify under received discrimination by proxy doctrine such as presented by, for instance, redlining. The fact alone that he is aware of an asymmetric distribution of an AI model's output does not entail a disparate treatment case.[169] Additionally, it would bring about a result which is somewhat paradoxical, especially in the case of for scoring credit default risk.[170] The better and more granular the model, the more likely it would be considered illegal. For understanding the third lender, this brings me to disparate impact doctrine and to the question when inequality of output makes a decision unlawful.

---

[166] See also United States v. Sineneng-Smith, 590 U.S. (2020), p. 19.
[167] But see the intense debate on what constitutes the counterfactual when discussing sex in the wake of *Bostock*, Berman/Krishnamurthi (2021); Dembroff/Kohler-Hausmann (2022), pp. 60 et seq.; Eidelson (2022), pp. 788, 794 et seq.; Koppelman (2022).
[168] See Adams-Prassl et al. (2022), pp. 14 et seq. for what they call the "perfect proxy".
[169] See above E.II.1; Adams-Prassl et al. (2022), p. 8 claim that this is a "startling conclusion". I disagree. Credit underwriting presents a good example: Lenders will often be aware of differences in credit default risk across protected groups. This can be a result of statistical discrimination, if protected groups, statistically, present a higher credit default risk see above C. Understanding this fact alone as troubling conflates statistical and taste-based discrimination.
[170] See above C for different definition of success, such as allowing to find especially vulnerable borrowers.

III. Disparate Impact

Disparate impact doctrine deals with proxies (facially neutral variables which correlate with a protected characteristic). To use proxies as building blocks for a decision is not prohibited *per se*. However, if this leads to a disproportionate distribution between protected communities and others, a justificatory reason must be established. (Only) in that limited sense does disparate impact doctrine follow an output-oriented logic, a point I will come back to.[171]

There is no comprehensive US Supreme Court guidance as to whether disparate impact doctrine applies to access to credit. *Ricci* seemed to limit the doctrine,[172] but in *Inclusive Communities* the Court held that "disparate-impact claims are cognizable under the Fair Housing Act (…) when their text refers to the consequences of the actions".[173] The ECOA lacks a results-oriented language of this type. Still, the CFPB and some courts seem open to applying disparate impact in that area.[174]

Under current EU law, anti-discrimination Directives require loans to be qualified as a "product or service offered to the general public". If this cannot be established, the outcome of the case varies across EU Member State laws on anti-discrimination. With the ongoing reform of the EU Consumer Credit Directive/2021, an explicit, broad anti-discrimination rule will be introduced into EU law.[175] Given the established terminology of previous anti-discrimination Directives, it is to be expected that this rule will cover both, direct and indirect discrimination.

1. Establishing a Disparate Impact Case

In the US, a disparate impact case can be made under the burden-shifting framework developed in Title VII cases. The plaintiff must make a prima facie showing of a disparate outcome for a protected group which includes identifying a facially neutral attribute. For an outcome to be disparate, a set of persons must be identified and the outcome for these persons must be compared to the rest of the population. If the relevant set is faced with a less favorable outcome, a prima facie case is established.[176] Along similar lines, the plaintiff has in the EU to establish facts which show a disparate output across groups, then the burden of proof shifts back to the defendant to show a justificatory reason.[177]

a) The benchmark of "similarly situated" persons

After establishing a prima facie case, the burden shifts to the defendant to demonstrate that there was a justification which explains the disparity. This could be done by discussing the benchmark for

[171] See below E.III.1.c.
[172] Ricci v. DeStefano, 557 U.S. 557 (2009).
[173] Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507 (2015), p. 2519.
[174] Ramirez v. Greenpoint Mortgage Funding, Inc., 633 F. Supp. 2d 922 (2008), pp. 926–927; Gillis (2022), p. 23.
[175] See above A.
[176] See Kim (2022), p. 17 on difficulties in practice to collect data about outcomes across the applicant pool.
[177] See ECJ ECLI:EU:C:1993:859 on burden of proof after a plaintiff has established statistical proof; see also ECJ ECLI:EU:C:2019:828 (note 56); ECLI:EU:C:2013:122 (note 42 et seq.).

comparison. Assume, for instance, that considerably more than 50 % of denied loan applicants are female. If (roughly) 50 % of the population are female, this looks like a disparate impact case. The lender can claim that this is not the right benchmark. Instead, he might suggest that only similarly situated sets of applicants ought to be compared. To decide which set is similarly situated to another set, he could propose to look at variables such as net worth, income, or credit history, all of which influence credit default risk. The effect might not be disproportionate if, for similarly situated sets of loan applicants, no sex discrimination shows. It is obvious that many cases will turn on building and comparing such sets of loan applicants. The narrower the group which serves as benchmark for a disparate impact comparison, the more difficult to establish a case.[178]

b) Justifying disparate outcome

How to justify disparate outcome varies significantly across context. A Title VII case of employment discrimination calls for different reasons than a housing or a credit underwriting case. If the discriminator has successfully demonstrated a business necessity defense along those lines, the burden shifts back to the plaintiff to show that there was a less discriminatory way to achieve that same goal.

The most natural justification for disparate outcome in a loan context are differences in credit default risk. Scoring and evaluating credit risk as such is an integral part of any credit decision.[179] The statistical discrimination this produces[180] does not automatically translate into discrimination under the law, even if this exercise produces disparate impact on protected groups. There are micro-arguments of the lender's business strategy to consider. [181] Then there are more general, macro-arguments which regulators and legislators take into account. They include preventing over-indebtedness of borrowers, allowing for risk-adjusted pricing across the population of borrowers, respecting freedom of contract and shareholder value goals of lenders, and preserving financial stability. This is not to say that these policy reasons automatically justify any unequal output produced, but to highlight the balance between competing interests.[182]

Strategic business goals of the lender which go beyond credit default risk, i.e. the ability to repay a loan with interest, pose more complicated questions. Personalized pricing of a loan based on its value for the specific borrower provides one example.[183] Borrowers in urgent need of a loan or with less sophisticated knowledge when evaluating and comparing interest rates will often be a vulnerable target for predatory lenders.[184] In some cases, this practice has been dubbed "reverse redlining".[185] In the US, some federal action has been taken through the Secure and Fair Enforcement for Mortgage

[178] Noting that there is little guidance on this question: Gillis (2022), p. 72.
[179] See above B.
[180] See above C.
[181] Hurlin et al. (2021) offer a methodology to distinguish whether a lender discriminates only for creditworthiness.
[182] See below F.IV.
[183] See below F.III.
[184] DeYoung/Philipps (2006).
[185] Fisher (2009) pp. 126 et seq.

Licensing Act[186] or the CFPB's qualified mortgage rule.[187] The DOJ, the CFPB and the OCC have in October 2021 announced the launch of a "Combatting Redlining Initiative" which includes "modern-day redlining" and discriminatory algorithms.[188] At the same time, strategic pricing is not illegal per se and the debate if it could provide a justification to offer less advantageous conditions to protected groups has only just begun.[189]

c) Disparate Impact and Output Control

Disparate impact doctrine, broadly speaking, has two distinct conceptual underpinnings. Some focus on input in the form of the facially neutral variable. If there is a clear correlation between the variable and the unequal outcome, the variable becomes "suspicious", as it were. For these theories, disparate impact is similar to "an evidentiary tool used to identify genuine, intentional discrimination – to "smoke out," as it were, disparate treatment".[190] Equality is understood as a formal (or anti-classificatory) concept.[191] By contrast, substantive (anti-subordinative, transformative) theories[192] are interested in "the consequences of [...] practices, not simply the motivation".[193] They understand unequal consequences as "disturbing in itself".[194] It is important to note that even though anti-subordinative theories look to consequences, they still require causation. Plaintiffs must show that a distinct building block, along a chain of causation, lead to the decision. Put differently: Disparate impact doctrine, in the form courts in the US[195] and in the EU[196] understand it today, does not provide for a "pure" output control. A pure output control would require plaintiffs solely to show that the decision produced inequality across groups. Instead, for a disparate impact case, just like for disparate treatment, causation between distinct variables and the decision needs to be established.[197] I will come back to this.[198]

---

[186] Fisher (2009), p. 153.
[187] O'Keefe (2016).
[188] https://www.justice.gov/opa/pr/justice-department-announces-new-initiative-combat-redlining (last accessed 27 October 2022).
[189] See in more detail below at F.III.
[190] Ricci v. DeStefano, 557 U.S. 557 (2009), discussed at Gillis (2022), p. 25 ("intent-based").
[191] Overview at Langenbucher (2020), p. 554; Sacksofsky (2017).
[192] Overview at Langenbucher (2020), p. 554; Sacksofsky (2017).
[193] Griggs v. Duke Power Co., 401 US 424 (1971), p. 432; Smith v. City of Jackson, 544 U.S. 228 (2005), p. 236; Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507 (2015), p. 2522; discussed at Gillis (2022), p. 26 ("effect-based"); see Mayson (2019), p. 2241 for understanding any output-control as aligning with anti-subordination.
[194] Sunstein (2019), p. 506.
[195] Where the Supreme Court has stressed the "consequences of actions", it has not established an output control of this type either. Rather, when speaking of "the consequences of actions" it was concerned with delineating disparate impact doctrine from a focus just on "the mindset of actors", see Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507 (2015), p. 10. See further below at E.III.3.a. on Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507 (2015) not allowing liability imposed solely based on a showing of a statistical disparity, id. p. 18 and stressing the importance of a „robust causality requirement" id., p. 20.
[196] See ECJ ECLI:EU:C:1993:859 on burden of proof after a plaintiff has established statistical proof; see also ECJ ECLI:EU:C:2019:828 (note 56); ECLI:EU:C:2013:122 (note 42 et seq.).
[197] Gillis (2022), p. 27.
[198] See below E.III.3.b.

2. The Third Hypothetical Lender: Opaque Bundles of Proxies

The third hypothetical lender I described above did not worry much about the variables or correlations in his underwriting model if the model was helpful in predicting credit default risk. He used a vast number of variables and a model which does not instruct its user about the precise correlations it had identified. Like Upstart, he uses a bundle of variables, such as education and employment history, which traditional lenders had not previously addressed. Against this backdrop, let us revisit Upstart and the CFPB's no action letter.

From what can be gathered based on publicly available information, the CFPB chose an unusual test to assess any disparate impact caused by Upstart's model.[199] Based on data provided by Upstart, it simulated outcomes under Upstart's proprietary model and compared them with outcomes under a hypothetical model using FICO scores. This simulation saw Upstart approving 27 % more borrowers than traditional lending models. Personal loan interest rates were 16 % lower on average.[200] It also found no disparities for minorities, females, or 62 years or older applicants.[201] The Bureau understood these findings as excluding disparate impact concerns. This stands in stark contrast to the mystery shopping exercise referenced above.[202] For Upstart's underwriting model, this report found significant differences between Black, Latinx and white persons as to both, loan origination fee and interest rate.

Arguably, one of the reasons for the disparity in outcome between the CFPB and the mystery shopping report has to do with the benchmarks used and the variables tested. The mystery shopping report held all other inputs constant and varied only educational background. This is how received doctrine and courts understand a disparate impact case:[203] The decision-maker (Upstart) is motivated by a motivational building block (educational background). It is facially neutral but leads to a disproportionately asymmetric effect (HSI penalty of $1,274 and HBCU penalty of $3,499 for a loan of $30,000). The logical next step would have been to explore potential whether this occurred among similarly situated persons. The authors of the report suggest this, given that they used the same college major, the same occupation, and the same annual income. Upstart might reply that having attended a differently ranked college makes these persons not similarly situated.

The problem with the mystery shopping exercise's methodology is that Upstart, just like the third hypothetical lender, uses a bundle of variables. At first glance, disparate impact doctrine has a ready answer for bundles of variables: they are a case of mixed motives.[204] Most courts and scholars frame

---

[199] Critical as to this method: Student Borrower Protection Center (2020), p. 21 fn. III but see the following text for a critique of the mystery shopping exercise.
[200] CFPB (2019); Upstart Blog: An Update from CFPB on Upstart's No-Action Letter, available at: https://www.upstart.com/blog/an-update-from-cfpb-on-upstarts-no-action-letter (last accessed 27 October 2022).
[201] CFPB (2019).
[202] See above C.
[203] See above E.III.1.
[204] See above E.II.2.a.

them as a problem of but-for causation. If (i) the variable was one among several building blocks along the chain of causation and (ii) removing it changes the outcome, causation is established.[205]

However, at closer inspection this reasoning does not capture how most AI and big data underwriting algorithms work. Test-prong (i) (the variable was one among several building blocks) is easily proven. However, test-prong (ii) (removing it changes the outcome) will only be met for models with limited data input. For these limited-input models, removing one variable (in Upstart's case: education) might change the outcome. By contrast, sophisticated AI models working with many input data are unlikely to meet test-prong (ii). With one variable eliminated, due to redundant encoding and the flexibility of multivariate regression, the underwriting model will fall back on other variables. Alone or in their combination, they can fill in for the removed variable.

In theory, one could go through many rounds of eliminating variables and fill-in variables. However, the more variables are eliminated, the less useful the model becomes because it loses predictive power.[206] Applying disparate impact's test-prong (ii) along those lines would penalize the more sophisticated algorithms. This hurts innovation, along with its inclusive potential, if lenders instead stick with less sophisticated or with traditional models. Additionally, there is a good chance that this strategy would not even encourage lenders to explore the discriminatory potential of their model. Depending on how much information a lender receives on the level of disparity which the regulator accepts, the lender's model might learn to produce a result which is just good enough. This might be a less costly strategy than engaging with input elements. All of this explains the prediction of scholars that, with improving technology, eliminating input variables will cease to be a promising regulatory avenue.[207] It also implies that one of anti-discrimination law's core elements, namely its leverage via distinct variables, will lose significance.

3. Why not understand the model as the building block?

If the received understanding of disparate impact doctrine runs into problems with sophisticated algorithms and bundles of proxies, why not understand *the model itself* as the facially neutral attribute, the "policy or practice" in the words of the US Supreme Court?[208] Along those lines, a disparate impact test would look as follows: (i) identifying the lender's use of the model as the facially neutral practice, (ii) showing an unequal outcome across groups and (iii) establishing causation.

---

[205] The mixed-motive test as described here has been developed in the context of disparate treatment cases, see Berman/Krishnamurthi (2021), pp. 99 et seq. If one were, instead, to argue that the but-for causation standard does not apply in disparate treatment cases, but, rather, a stricter causation standard is in order, where the suspicious variable was the *sole* motive, the mystery shopping exercise would be (even more) inappropriate: Upstart uses a bundle of variables. Put differently: education was not the sole motivational building block for Upstart's model.

[206] Hellman (2020), pp. 830, 836 on the loss of confidence in the information provided by the algorithm and why this matters in different ways, depending on normative context; Mayson (2019), p. 2249: "imposing certain metrics of output equality will therefore have a cost in accuracy"; Sunstein (2019), p. 509: "tradeoff between accuracy and fairness".

[207] Fuster et al. (2022), p. 8; Gillis (2022), pp. 47 et seq.

[208] Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507 (2015).

There are two concerns with this understanding. The first has to do with the counterfactual when establishing causation. The second addresses the causation requirement as such.

a) Does the model cause disparate impact?

Received doctrine requires the plaintiff to establish causation when making a disparate impact case, this has surfaced throughout the paper. Causation as a test-prong for disparate impact is not a fault-based inquiry in that the discriminator caused the underlying inequality. US examples concern, for instance, a requirement of high school diplomas in an employment context[209] or an IQ test in deciding on special education classes.[210] The Court decided that the causation requirement was met upon proof that the high school diploma or the IQ test disproportionately impacted African Americans. It was not interested in exploring whether the discriminator was aware of his practice's discriminatory potential. Even less did it discuss who was at fault for African American children having high school diplomas at lower rates or gaining a lower score at an IQ test than white Americans.

At the same time, the US Supreme Court has ruled out a showing of statistical disparity as the only indicator for a disparate impact case. Proving statistical disparity only, the Court held, would be incompatible with the required "robust" proof of causation, meant to ensure that defendants are not "held liable for racial disparities they did not create".[211] A "robust" proof has to show that statistical significance is "sufficiently substantial"[212]. This rules out situations where random sampling caused the disparity.[213] It also rules out cases where the same disparate impact would have existed without the defendant engaging in the challenged practice.[214] It is this latter thought which renders the causation element tricky in the area of credit underwriting.

What will proof of causation look like if we understand the AI model as the facially neutral policy or practice? The case will then turn on what the hypothetical counterfactual is. Maybe the AI model was just one out of many elements of the credit decision which was primarily taken by human credit officers. This would allow to investigate the AI model's impact. If the human credit officer had come to the same conclusion, with or without the model, the plaintiff cannot establish causation. The hypothetical counterfactual would trigger the same result. If the human credit officer would have come to a different conclusion without the model, the plaintiff can show causation.

By contrast, if the human credit officer always followed the model's recommendation or, even more, if the underwriting process was entirely automated, it is not clear how to show causation. Removing the AI model from the chain of causation leaves us without a guideline for assessing the

[209] Griggs v. Duke Power Co., 401 US 424 (1971), p. 431.
[210] Larry P. by Lucille P. v. Riles, 793 F.2d 969 (1948), p. 983.
[211] Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507 (2015), p. 2523.
[212] Watson v. Fort Worth Bank & Trust, 487 U.S. 977 (1988), p. 995.
[213] Groves v. Alabama State Bd. of Educ., 776 F.Supp. 1518 (1991), pp. 1527–1528.
[214] Ellston v. Talladega Cty. Bd. of Ed., 997 F.2d 1394 (1993), p. 1415.

counterfactual. In that respect, the situation differs from the (traditional) focus on distinct building blocks. Above, I explored an AI model's individual data points, understood as building blocks along a chain of causation.[215] The worry was that this worked well only for limited-input models. By contrast, the higher the number of variables and bundles of proxies, the more likely that removing a single data point would not change the outcome. Theoretically, it was possible to establish the hypothetical counterfactual by running many rounds of removing single variables and exploring whether there was a change in outcome. However, requiring lenders to downsize sophisticated algorithms in this way seemed less useful as it penalized the more powerful algorithms. Now, the focus is not any more on individual data points, but on the entire model. Removing it without a human credit officer giving additional input, means losing all criteria for decision-making.

In the face of the lack of a counterfactual, it is interesting, once again, to return to the CFPB's no-action letter for Upstart. The Bureau used a novel type of simulation test. It did not follow the traditional routine of pointing out one variable, establishing disparate impact and looking for justificatory reasons. Instead, the CFPB measured the number of persons who would be eligible for a loan under a hypothetical FICO-score model against the number of persons eligible under Upstart's model. The Bureau repeated the exercise comparing interest rates and access to loans for protected communities. For all these tests it found that borrowers fared better under Upstart's model than under a (hypothetical) traditional model. As to access to credit, this was true for the absolute number of borrowers across all groups as well as for the absolute number of protected-group borrowers. Put differently: Minority borrowers had better chances to be eligible for a loan under Upstart's model than under the hypothetical traditional model. However, if one zoomed in on one subset of minority borrowers (Black and Black-Hispanic persons), the distribution was still skewed. Minority borrowers which were eligible under Upstart's model were facing disadvantages when compared with the subset of white and white-Hispanic persons eligible under Upstart's model. This was true as to relative numbers of access to credit, origination fees and interest rates. It is also consistent with the empirical studies described above.[216]

It is hard to say whether one should applaud the CFPB for this unusual strategy or criticize it for overreaching its authority. The Bureau's thinking might have been: If in absolute numbers more protected-group-borrowers have access to loans than under a hypothetical FICO score, this provides for more inclusion. Against that background, the Bureau might have claimed, it does not matter if the surplus is unequally distributed: everyone is better off.[217] But is this a convincing argument? Looking to the hypothetical simulation with FICO as the counterfactual is not entirely unreasonable in a country where access to loans follows a standardized routine. It rewards lenders who offer an advantage, at least for some groups and at least if compared with the current situation. Two downsides are apparent. The current FICO-based standard remains the benchmark. This can hurt innovation and it can defeat the purpose of those FinTechs who wish to offer access to credit for borrowers who do not perform well under the traditional metric. Additionally, the CFPB's aggregate-view test is hardly compatible with received disparate impact doctrine. Anti-discrimination law is about relative

[215] See above E.III.
[216] See above E.III.2.
[217] Langenbucher/Corcoran (2022), p. 156.

disadvantages of one group when compared to another group. The Bureau focused instead on the surplus produced by Upstart's model, irrespective of the relative composition of the group of borrowers. This is not to say that such criteria are irrelevant or should not be pursued. They just cannot be justified as an application of received anti-discrimination law. Instead, they point towards a quality control of the model.[218] Quality controls focus directly on the model, without taking a detour, as it were. In that way, they do not need to supplement disparate impact's causation test with a novel type of simulation test.

In its 2020 renewal of the no-action letter, the CFPB even more clearly reorients its investigation away from disparate impact doctrine and towards an investigation into the model. It does not ask for causation or try to establish a hypothetical counterfactual to understand what would happen if the model were removed. Rather, the CFPB mentions a Model Risk Assessment Plan which Upstart is required to follow.[219] This includes model documentation as well as monitoring how Upstart's customer population and model performance change over time. The CFPB explicitly asks for "access-to-credit testing". Additionally, the Bureau mentions testing the "model and/or variables or groups of variables" for disparate impact and predictive accuracy by group as well as "research approaches that may produce less discriminatory alternative algorithms that meet legitimate business needs". A similar approach is suggested by the EU AI Act and I will come back to model quality checks along those lines further below.[220]

b) Do we need a causation requirement when testing AI models?

Before following the CFPB's path beyond received anti-discrimination doctrine, one last comment on causation is in order. There are good reasons for received doctrine to require a causation element in disparate impact cases. One of them is to single out attributes or practices which courts suspect of deliberately hiding discriminatory intentions if disparate treatment by proxy cannot be established.[221] Another one is to incentivize defendants who are aware of their model's discriminatory potential to look for alternatives which are less discriminatory. Yet another one has to do with limiting responsibility. Disparate impact is not a fault-based responsibility. At the same time, a defendant is not responsible for a wrong "he did not create".[222] A causation element is one way to link actions of the defendant to the discriminatory outcome.

Much of this is modelled on a human actor who hides discriminatory intentions, is aware of discriminatory potential or unwilling to search for alternatives with less discriminatory potential. Machine-learning algorithms currently used in credit underwriting lack an analogue to intentions,

[218] See below F.I.2.
[219] CFPB (2020a).
[220] See below F.I.
[221] See the "smoking out disparate treatment" argument in Ricci v. DeStefano, 557 U.S. 557 (2009).
[222] Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507 (2015), p. 2523 on racial disparities.

awareness, or unwillingness.[223] Instead of causation, algorithms look to correlation.[224] This might be one of the deeper reasons for the ill fit not only of disparate treatment, but also of disparate impact doctrine for handling algorithmic discrimination. One reaction is to tune down the importance of the causation element in disparate impact cases when dealing with AI. This paper suggests a different path for future research. First, it points towards model control as a core requirement for well-trained models. Second, it invites to explore targeted interventions where decisions from a pre-AI time might need adjusting or even a normative double-check. I conclude with outlining contours of controls and double-checks along those lines and hope to provide details in follow-up papers.

F. NEXT STEPS: TOWARDS A REGULATORY DESIGN FOR CONSUMER CREDIT IN THE AGE OF AI

The CFPB has not published data on the reasons which explain the asymmetry in origination fees and pricing across protected groups in Upstart's underwriting model. As described above, there are many possible explanations.[225] Disproportionate effects can go back to differences in credit default risk which the model found, representing a risk premium and reflecting existing inequalities in the world. They can also be triggered by bias in the data or by bias in the model.[226] In this case, they show an inadequate understanding of the credit default risk of some groups. Alternatively, Upstart's business model might provide the explanation. Maybe, it developed a profile which works especially well for white persons with an educational background in a predominantly white college who do not perform well under standard FICO-metrics. This could explain the surplus in borrowers the CFPB found with its FICO-simulation. Yet another reason could be strategic pricing. Upstart's model could have figured out that the probability to accept less favorable terms was higher in protected communities. Its optimization goal (or: definition of success)[227] is to blame if the model consistently offered higher origination fees and interest rates to these groups.

Some of these explanations call for a regulatory framework requiring careful examination of model and data.[228] Biased models can produce inappropriate risk assessments or leeway for inefficient rent-seeking activities.[229] Additionally, these explanations suggest that there is a strong case to be made for clear rules and efficient enforcement in the area of credit reporting, scoring and financial privacy.[230] Other explanations suggest the normative double-check I mentioned earlier,[231] especially

---

[223] See Wachter (2022), pp. 31-32 on explaining that "discriminatory behaviors carry with them an assumption of moral superiority" and claiming that "these notions make a lot of sense from a human lens (…) however, when algorithms being used it is not fully clear if their grouping invokes the same moral wrong".
[224] Burk (2021), p. 1147.
[225] See above E.II.3. and E.III.2.
[226] See Gillis (2020), p. 18.
[227] O'Neil (2016), p. 21, see below F.III.2.
[228] See below F.I.
[229] Aggarwal (2021), pp. 50 et seq.
[230] See below F.II.
[231] See above E.III.3.b.

if personalized prices hurt vulnerable applicants.[232] Then, there is the question who bears the cost of offering equal access to persons who are not similarly situated.[233]

I. Quality and Governance Control

Quality issues of the AI model or the data can hurt both sides of the transaction. The borrower pays too much interest or is not eligible for a loan. The lender leaves money on the table if he denies a loan because the model incorrectly sends the applicant to a statistical bucket, he does not belong in. The EU AI Act[234] tries to respond to this by treating AI models as products in need of regulation.[235] The Act requires risk management systems which include continuous processes and regular updating.[236] Data governance and management practices look at training, validation and testing data,[237] models must be regularly re-trained, and human oversight by "natural persons who have the necessary competence, training and authority"[238] must be ensured. Along similar lines, authors such as Citron and Pasquale suggest licensing and audit requirements,[239] the FTC has initiated an advanced notice of proposed rulemaking and the US AI Bill of Rights talks about safe and effective systems.[240]

1. By way of illustration: A brief glance at the EU AI Act

The AI Act introduces a risk-based approach for AI "systems".[241] A small number of AI systems are impermissible. Many face minimal or no compliance requirements and some are considered high risk. AI underwriting and scoring models fall in the high-risk category. The reason is that these systems "determine (…) access to financial resources" and their use "may lead to discrimination of persons or groups and perpetuate historical patterns of discrimination (…) or create new forms of discriminatory impacts".[242]

For high-risk systems, the Act follows the logic of regulating dangerous products, similar to the US Blueprint for an AI Bill of rights which lists the need for "safe and effective systems".[243] It (roughly) distinguishes five categories of compliance requirements which focus on data and data governance, technical documentation and record-keeping, transparency, human oversight, and checks on robustness, accuracy and cybersecurity. All these rules concern professional developers and users

[232] See below F.III.
[233] See below F.IV.
[234] Proposal for a Regulation of the European Parliament and the Council Laying down harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) of 21 April 2021, COM(2021) 206 final. The text is based on the 4th Presidency Compromise Text of 10 October 2022. For better readability, I refer to this text as: AI Act.
[235] See Langenbucher (2022); Langenbucher/Corcoran (2020).
[236] Art. 9 AI Act.
[237] Art. 10 AI Act.
[238] Art. 29 para. (1a) AI Act.
[239] Citron/Pasquale (2014), p. 21 for employment, insurance and health care; id., pp. 24 et seq. on the FTC's statutory authority to combat unfair trade practices as to scoring.
[240] FTC (2022).
[241] Art. 3 para. (1) AI Act.
[242] Recital (37) AI Act.
[243] Available at https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (last access 27 October 2022).

only. By contrast, the Act does not address the situation of end consumers but, so far, delegates it to the private law of the EU Member States. Looking ahead, the EU Consumer Credit Directive/2021 proposes an anti-discrimination rule.[244] Additionally, the reform of the EU Product Liability Directive takes up some concerns of liability for AI systems.

Training, validation, and testing data sets are to undergo data governance checks. The drafters aim at data which is "to the best extent possible free of errors and complete".[245] They are aware of the pitfalls of collecting alternative data and ask developers to evaluate "availability, quantity and suitability of data sets" [246]. Developers must make sure their data set has the "appropriate statistical properties"[247] and reflects specifics of "the geographical, behavioral or functional setting within which the high-risk AI system is intended to be used"[248]. Additionally, they must identify "data gaps or shortcomings"[249] as well as "possible biases (…) that lead to discrimination prohibited by Union law".[250] To detect these, the proposal permits processing of sensitive data on protected characteristics if this is "strictly necessary for the purpose of bias monitoring".[251] The Act requires "appropriate safeguards (…) including technical limitations on the re-use and use of state-of-the-art security and privacy-preserving measures, such as pseudonymisation, or encryption" [252].

Developers must run compliance checks on their AI systems before putting them on the market. If they provide high-risk AI systems they must ensure that these systems undergo the relevant conformity assessment procedure and draw up an EU declaration of conformity.[253] Following the EU's new legislative framework (NLF),[254] the developers of the product carry out these conformity assessments.[255] For some products this involves a conformity assessment body, a private entity which Member States designate to run conformity assessments.[256] Developers of AI systems which are financial institutions under Union financial services legislation already follow a special compliance regime of regulated industries. A financial institution which uses a high-risk AI system fulfills various

[244] Art. 6 Proposal for a Directive of the European Parliament and of the Council on consumer credits, COM(2021) 347 final.
[245] Art. 10 para. (3), recital (44) AI Act.
[246] Art. 10 para. (2) AI Act.
[247] Art. 10 para. (3) AI Act.
[248] Artt. 10 para. (4), 13 para. (3) lit. b no. i.
[249] Art. 10 para. (2) lit. g AI Act.
[250] Art. 10 para. (2) lit. f AI Act.
[251] Art. 10 para. (5) AI Act.
[252] Art. 10 para. (5) AI Act.
[253] Artt. 43 et seq. AI Act.
[254] See Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011, 2019 O.J. L (169) 1; Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93, 2008 O.J. L (218) 30.
[255] For market surveillance see: Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20. June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011, 2019 O.J. L (169) 1.
[256] See https://ec.europa.eu/growth/tools-databases/nando/ (last access 27 October 2022).

monitoring obligations of the AI Act by complying with the relevant Union financial services legislation.[257]

To ensure adequate market surveillance, Member States designate national competent authorities.[258] The AI Act's institutional design rests on various regulatory agencies, depending on the entity which uses the AI model or puts it on the market. As an integral part of financial services oversight, the competent regulator will supervise compliance with the AI Act if there is "a direct connection with the provision of those financial services"[259] and unless the Member State has identified another relevant authority.[260] However, not all entities which are involved in scoring or in extending credit of some sort are financial institutions. Many Fintech platforms are not financial institutions,[261] nor are scoring agencies, insurance companies[262] or companies which the AI Act refers to as offering "essential private services" such as housing, electricity, and telecommunication. These non-banks fall under the jurisdiction of newly to be established regulatory agencies, entrusted with monitoring use and development of high-risk AI. Additionally, AI regulatory sandboxes are geared towards promoting innovation.[263]

The AI Act does not address private litigation.[264] In line with its spirit of product regulation, it speaks to developers and professional users, not to retail consumers or borrowers. A right to complain to market surveillance authorities is included,[265] but private rights of action for damages of retail borrowers fall under EU and Member State law. However, the proposal for an AI Liability Directive[266] takes up some of these claims. While the plaintiff must establish defectiveness of the product, the damage suffered and causation, the Directive includes a presumption of defectiveness and of causation in certain situations. The drafters of the reform start from the assumption that fault-based liability, when faced with the complexity, autonomy and opacity of AI, makes it impossibly hard for plaintiffs to establish a case. For non-contractual liability, the proposal shifts the burden of proof to the defendant and includes discovery provisions for plaintiffs seeking damages.

2. Controlling Quality of Data and Models

An AI scoring outfit that uses historically biased training data will often come up with underwriting models which present a snapshot of reality at some point in time – what I have called "yesterday's world".[267] Its value for assessing borrowers depends on the match between the snapshot and today's

[257] Artt. 17 para. (3), 18 para. (2), 20 para. (2), 29 para. (4) subpara. (2), para. (5) subpara. (2) AI Act.
[258] Art. 59 AI Act.
[259] Art. 63 para. (4) subpara. (1) AI Act.
[260] Art. 64 para. (4) subpara. (2) AI Act.
[261] Art. 37 EU Consumer Credit Directive/2021 requires Member States to ensure that Fintech platforms which match lenders and borrowers ("credit intermediaries" and "providers of crowdfunding services") fall under licensing and supervision by an independent competent authority.
[262] See recital (37), Annex III para. 5 lit. e.
[263] Artt. 53 et seq. AI Act.
[264] See recital (5a).
[265] Art. 63 para. (11) AI Act.
[266] Proposal for a Diretive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final.
[267] See above D.I. and Hurlin et al. (2021), p. 3 on using training data based on past decisions made by biased loan officers.

world. If the way in which today's world is different from yesterday's world does not matter for credit default risk, using old data might not hurt much. But if the training data implies, for instance, restrictions on taking out loans or an income distribution across sex or race which do not correctly represent today's world, the model will come up with skewed scores. The same is true if the score attributed to applicants who share features of historically privileged communities is higher than what their actual situation suggests.

Against this background, there seems to be a straightforward solution: Why not run compatibility checks between yesterday's and today's world? Unfortunately, a core problem for controlling for biased data is not only about identifying the extent of that match. In many cases, the lack of compatibility is not apparent, or the problem extends to choosing variables. An often-cited example for this latter concern has to do with a decision-making algorithm used by US hospitals. The algorithm allocated patients to programs improving care for those with complex medical needs.[268] A machine-learning (ML) research team at UC Berkeley received data from a hospital to work on ML and health care services. The researchers were surprised to find lower risk scores for Black persons which were equally sick as white persons who were assigned higher scores. They found that the algorithm looked to total health-care costs per year as the variable to assess risk scores. However, as an indication of how sick a person is, this proved wrong for Black persons. Health care administered to them cost considerably less per year than health care provided to a white person with a similar health profile. As a result, Black persons had to be much sicker than white persons to be allocated the same risk score. Put differently: They had to wait much longer to receive the personalized care for patients with complex medical needs.

The example illustrates the plethora of problems. One problem has to do with awareness of data or model quality. Without the researchers and their statistics, the problem with the model's choice of variable might not have been detected at all. A possible fix for this are regular model audits which include a thorough examination of the underlying variables and assumptions.[269] Variables can be fitted to subgroups, if the traits "are more predictive for one race than for another", as Hellman suggests.[270] This strategy can help to cope with the example discussed earlier where non-performance was used as a core variable but implied different things across groups.[271] If a model attributes equal weight to non-performance across groups of borrowers, this might not adequately reflect credit default risk of each person in the population. If one group of the population consistently faces higher interest rates despite being similarly situated, the probability of default in that group will be higher.[272] But this is not because of an initially lower creditworthiness of each member of this group, but because of the higher burden to pay back. To raise awareness in those and similar situations, diverse coding teams and awareness trainings have been suggested as a step towards better model quality

[268] Obermeyer et al. (2019), on this example Burrell/Fourcade (2021); Langenbucher (2020), p. 555; Ledford (2019).
[269] Ledford (2019); see the "equity assessments" which the US Blueprint for an AI Bill of Rights suggests under Algorithmic Discrimination Protections, available at https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (last access 27 October 2022).
[270] See Hellman (2020), pp. 853 et seq., providing examples which have been discussed in the literature.
[271] See above D.I.
[272] See O'Neill (2016), p. 144: "nasty feedback loop".

control.[273] The AI Act requires that "high-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations ('feedback loops')".[274]

Still, there are no easy and straightforward fixes. Better variables to predict risk might be hard to find. The users of the model might prefer to go ahead with an imprecise model, rather than with no model at all. This is especially true if the imprecise model still performs better than biased and cognitively limited human credit officers.[275]

Additionally, the example of the hospital algorithm concerns a conscious choice of variable (total health care costs per year) by the coders. If ML algorithms take over the process of choosing variables and attributing weight to them, quality checks are much more challenging.[276] Reverse engineering to identify a set of core variables has been proposed as a model quality check. Competition between AI models is another approach. Data audits can help if some groups are penalized because of a lack of relevant data on the members of this group. None of these strategies will necessarily help with historical bias when conditions change.[277] Take, for instance, a rule under which married women were not eligible for credit unless their husband signed the loan contract. It is one thing to have the model integrate the change between yesterday's and today's world, once legal reform allows unmarried women to sign a loan on their own. But (like the opaque bundles of proxies discussed above) the status as a married woman will be encoded redundantly in many other variables. Tweaking the model by fitting it to the subgroup of unmarried women will, for some time, make the model less precise for lack of data on that subgroup.

II. Credit Reporting and Financial Privacy

Information gathered from sources such as social networks, internet usage or behavioral tests is considerably more prone to mistakes and misunderstandings than traditional credit reporting data.[278] The US has no Federal law in place which specifically targets big data aggregators. However, the FCRA provides safeguards for borrowers who want to dispute completeness or accuracy of a credit report.[279] Section 1033 of the Dodd-Frank Act transferred rulemaking responsibilities under FCRA to the CFPB. Enforcement authority rests with the FTC. The EU AI Act expects data which providers

[273] Benjamin (2019).
[274] Art. 15 para. (3) subpara. (3) AI Act.
[275] Ledford (2019); reaching the same conclusion for the criminal justice system: Mayson (2019), p. 2277; however, see Kim (2022), p. 5 in the context of the inability of AI underwriting models to learn from false negatives (see D.III above): "one of the claimed benefits of AI – its ability to learn over time – is far more limited when used to make decisions about people".
[276] Citron/Pasquale (2014), p. 5.
[277] See above D.
[278] See above D.III.
[279] 15 U.S.C. § 1681(i)(a)(1)(A); Langenbucher/Corcoran (2022), p. 162.

use to be "relevant, representative, free of errors and complete".[280] Arguably, this sets a standard that will not only be hard, but at times impossible to meet.

1. Data aggregators

The US has since the 1970s had a statutory framework for credit reporting and scoring in place. However, many data aggregators which collect and process the vast input of non-financial information currently operate outside that regulatory perimeter. To qualify as a consumer reporting agency under the FCRA, the person must regularly engage in the practice of assembling and evaluating consumer reports.[281] The FTC has understood evaluation as "appraising, assessing, determining or making a judgment on such information". "An entity that performs only mechanical tasks in connection with transmitting consumer information is not a consumer reporting agency".[282] Instead, persons which perform mechanical tasks of this type are qualified as a "conduit" only.[283] Various data aggregators have claimed that they should be understood as a conduit, with FinTech lending platforms doing the evaluation of the consumer report.[284] The CFPB seems more open to bringing data aggregators under its jurisdiction. Section 1033 of the Dodd-Frank-Act is cited towards that end[285] and competitors of Fintech firms from the banking industry are urging the Bureau to do so.[286] The same goes for AI scoring agencies. They will not necessarily qualify as a consumer reporting agency under the FCRA nor will their AI-based scores always be considered a consumer report. A Fintech lender who builds a score for his own purposes using proprietary data falls outside the scope of the FCRA entirely, unless he furnishes data to third parties.[287] Additionally, most of FCRA's requirements are procedural and impose few limits on collecting data.[288] All of this suggests that the FCRA reflects the time and technology Congress thought it was responding to when passing the law. While the FCRA's policy goals remain valid, the advent of AI underwriting models and big data raises doubts whether the established regulatory design will keep pace with new developments, suggesting an overhaul of the statutory rules.

In the EU, the GDPR is the first comprehensive piece of legislation to formulate legal principles not only across Member States, but also across various situations which raise data privacy concerns. It follows an omnibus, rather than a sectoral approach and does not contain rules which zoom in on credit underwriting. Collecting, processing and furnishing data to third parties is only lawful

[280] See above F.I.1. and Art. 51 para. (3) AI Act.
[281] Kim/Hanson (2016), pp. 21 et seq.
[282] FTC (2011) p. 29; see for a narrow reading of the LexisNexis product "Accurint" which was not considered delivering "credit reports": Kim/Hanson (2016), p. 28.
[283] Id., p. 29.
[284] NCLC (2020), p. 8; see the Ninth Circuit on a similar argument when it decided that *Fannie Mae* was not a consumer reporting agency, Zabriskie v. Federal National Mortgage Association 912 F.3d 1192 (9th Cir. 2019); but see Kim/Hanson (2016), pp. 30 et seq. for other courts reaching a different conclusion.
[285] CFPB (2020b), pp. 71009 et seq.
[286] ABA (2022).
[287] Kim/Hanson (2016), p. 26; Langenbucher (2020), p. 535.
[288] Kim/Hanson (2016), p. 32.

according to a list of justificatory reasons.[289] Data aggregators and Fintech platforms will usually qualify as "data processors".[290] For data processing to be legitimate it must fall under one of the GDPR's exemptions. "Consent" seems a most natural one. However, the standard for "freely given" consent is strict. Additionally, consent is under the GDPR revocable. Hence, most data aggregators look elsewhere. Further exemptions cover data processing understood as a step taken at the request of the data subject prior to entering into a contract,[291] or data processing being lawful based on legitimate interests.[292] Should the non-traditional data involve protected categories, the GDPR additionally asks for "explicit" consent.[293] Under the EU Digital Markets Act (DMA)[294] data aggregators can not necessarily rely on consent if they qualify as a "gatekeeper".[295] This new restriction for gatekeepers covers the combination of personal data which the gatekeeper obtains as part of its core service with data from other platform services it provides. It also rules out "legitimate interest" as a justification for combining or cross-using data.

The reform of the EU Consumer Credit Directive/2021 directly targets credit underwriting. It starts from the assumption that the data lenders look to when checking credit default risk should be financial data "on the consumer's income and expenses and other financial and economic circumstances".[296] If the lender uses profiling and automated processing of data, he must offer human intervention, allow the applicant to contest the decision and require the lender to provide an explanation.[297] The use of alternative data is not unlawful, however, it needs to be "necessary and proportionate" and speak to financial commitments. Arguably, a more general behavioral evaluation of the applicant's character does not seem to be a lawful goal of credit default risk checks. Recital (47) declares that "personal data such as personal data found on social media platforms or health data, including cancer data (…) should not be used". At the same time, there is no explicit prohibition in the proposed Directive nor is it clear which standard will apply to data and creditworthiness evaluation furnished by third parties. Arguably, given redundant encoding and flexibility of multivariate regressions, it is doubtful whether an input control along the lines of the proposal is a promising avenue. The proposal seems somewhat undecided between neither prohibiting alternative data, nor encouraging its use, due to what seems are moral concerns. At the same time, it refrains from addressing the underlying tension between allowing for more granular predictions at the expense of sensitive personal data.

2. Accessing, verifying, and rectifying data

[289] In more detail Langenbucher (2020), pp. 534 et seq.
[290] Langenbucher/Corcoran (2022), p. 149, on FCRA id., p. 150, and Langenbucher (2020), p. 534.
[291] Art. 6 para. (1) lit. b GDPR.
[292] Art. 6 para. (1) lit. f GDPR.
[293] Art. 9 GDPR.
[294] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), 2022 O.J. L (265) 1.
[295] Art. 2 para. (1), (3), (5) DMA.
[296] Art. 18 para. (2) EU Consumer Credit Directive/2021.
[297] Art. 18 para. (6) EU Consumer Credit Directive/2021.

In the US, the FCRA regulates the type of information credit reporting and scoring agencies use and the rights of consumers in relation to credit reports. The statute strikes a balance between the interests of lenders in statistical discrimination and the consumers' right to financial privacy and accuracy of information.[298] Traditional consumer reporting agencies such as credit bureaus use a diligent process of collecting the financial data which inform their algorithms and their assessment of individual borrowers. Backing this up, the FCRA provides borrowers with rights to access information and to rectify incorrect entries.[299] In case of concerns, the borrower can notify the FTC which will conduct a reasonable reinvestigation to determine the accuracy of the information and, if incorrect, have it deleted.[300] Today, these interests remain unchanged. However, consumer rights to dispute the accuracy of information might not apply to data aggregators.[301] Many US states have started to introduce privacy laws or are in the process of doing so. The FTC has published an advanced notice of proposed rulemaking,[302] and the US Blueprint for an AI Bill of Rights includes a right to privacy.[303]

The EU GDPR requires any data controller to inform the person from whom it collects data and to explain how to get access.[304] Lenders, credit reporting agencies or scoring agencies who use automated decision-making or profiling qualify as data controller. Applicants can request confirmation on whether an entity processes personal data on him and what the purpose of such processing is.[305] They also have rights to rectification and erasure.[306]

In the face of big data mining, rights which provide for disclosure and verification are core tools. They further consumer protection. At the same time, they contribute to well-trained algorithms based on verified data. However, even if data collection or furnishing data to third parties is prohibited or if there is a claim to rectify incorrect data, enforcement is by no means straightforward under either, US or EU law.[307] Novel data aggregators might not be known to the general public. If they are, it might not be clear which type of data they collect, process, or furnish to third parties. Often, consent will be uninformed and without opt out options.[308] This can leave many consumers with a largely useless claim, lacking information on the identity of the party liable, facing consent they had no choice but to give, or not realizing the ultimate purpose of the data collection. Ideally, further work on the GDPR and the FCRA would provide for more stringent enforcement options. Until such work is done, one default option is regulatory action based on general rules and principles which authorize

[298] Langenbucher (2020), p. 534.
[299] See 15 U.S.C. § 1681i § 611 (a)(1)(A).
[300] Citron/Pasquale (2014), pp. 14 et seq.; Langenbucher (2020), p. 535.
[301] Above F.II.1. Citron/Pasquale (2014), p. 20; Langenbucher/Corcoran (2022), p. 151.
[302] FTC (2022).
[303] Available at https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (last access 27 October 2022).
[304] Art. 13 GDPR regulates data the entity collects, Art. 14 GDPR includes data received from third parties, Langenbucher (2020), pp. 539 et seq.
[305] Art. 15 GDPR.
[306] Artt. 16, 17 GDPR.
[307] Langenbucher (2020), pp. 535–536.
[308] Awrey/Macey (2022); FTC (2022) pp. 51274 et seq.; see Art. 34 EU EU Consumer Credit Directive/2021 on requiring Member States to promote financial education.

enforcement. Another one are rights of action for collective bodies charged with monitoring compliance of data collectors.

III. Transparency, Scoring, Optimization Goals and Responsible Lending

Today we can only speculate which sets of applicants will in the long-term profit from AI underwriting models. Most empirical studies I described in this paper suggest that there is a chance to expect more accurate information for applicants which are difficult to score along standard lines.[309] Some of these will be invisible primes, and AI underwriting models will allow better access to credit. Arguably, the added value for traditionally mainstream applicants will vary considerably according to the type of data, the type of AI model and the goal the model is optimizing. We might eventually be looking at outcomes which do not reflect what the current legal framework of anti-discrimination laws has in mind.[310] Consider, for instance, AI underwriting models which lead to unanticipated imbalances in the distribution of loan applicants outside the definition of protected groups: persons who do not take care of regularly updating their software, who do not have a social media presence, whose IoT fridge often signals a lack of alcoholic beverages or who regularly google information on moving houses. They might find it hard to get a loan approved, irrespective of their sex, race or religion. Black-box algorithms might make it hard or altogether impossible to establish which one of these variables drive an AI's score.

1. Commercial surveillance and Scoring

The interests of the lender who uses big data and algorithmic decision-making will typically remain unchanged. He looks for cost-effective tools to monitor loan applicants according to his business strategy. Possible strategies include pricing credit in tune with market standards as well as selling predatory loans. Lenders will, under both strategies, applaud AI scoring as presumably more objective, more efficient and more tuned to the individual person.[311]

By contrast, applicants find themselves in a radically different situation. They do not necessarily know which variables the AI underwriting model is looking for nor which weight is accorded to individual attributes. They might worry about a "world of conformity"[312] where consumers "fear to express their individual personality online" and constantly consider their digital footprints.[313] If the novel distribution does not reflect traditionally protected characteristics, anti-discrimination laws fail to protect the newly unsuccessful set.

[309] See above B.
[310] See Wachter (2022), pp. 15-29.
[311] Kim (2022), p. 1.
[312] Berg et al. (2018), p. 26.
[313] Berg et al. (2018), p. 6; Burk (2021).

In a world where humans are ignorant of the variables driving the AI underwriting model, additional concerns come to the fore. Scored persons might feel that they are exposed to arbitrary decisions which they do not understand and which are unexplainable even to the person using the algorithm.[314] Consumers might try to randomly change their online behavior in the hope for a better score.[315] Manipulation along those lines will work better for some variables (such as regularly charging a mobile device) than for others (such as changing mobile phone brand or refraining from impulse shopping).[316] One strategy is to mimic the profile of an attractive borrower. If this is costless, Berg and his co-authors submit, an uninformative pooling equilibrium evolves. All senders choose the same signal which does not help to reduce the information asymmetry between borrower and lender.[317] Those same authors suggest that firm behavior might adapt as well.[318] A firm whose products signal low creditworthiness could try to conceal its products' digital footprint. Commercial services may develop, offering such concealing services or making consumers' digital footprint look better. Along similar lines, the CFPB fears that the chances to change credit standing through behavior may become a random exercise.[319] This is even more worrisome if the optimization goals of algorithms include not only predictions on the best credit default risk but also on the candidate most likely to accept rates above market prices.[320]

Transparency is the natural remedy when faced with opaque decision-making. The GDPR illustrates this by providing a right to receive "meaningful information about the logic involved, as well as the significance and the envisaged consequence of such processing" if automated processing and profiling is intended.[321] Additionally, the data subject has a right to know what the purpose of processing of his data is.[322] The EU Consumer Credit Directive/2021 provides another illustration, requiring human intervention, an explanation of logic and risks involved in automated processing and ways for the borrower to express his view and contest his creditworthiness assessment.[323]

At the same time, achieving meaningful transparency is a complex endeavor. One concern is practical and has to do with the way in which transparent information is given to the consumer. Lines of code will for most people not help nor will statements about the lender using "all available data".

[314] CFPB (2017), p. 18; Burrell/Fourcade (2021), p. 226.
[315] Berg et al. (2018), p. 25 referencing the Lucas Critique, see Lucas (1976).
[316] Berg et al. (2018), pp. 25–26.
[317] Berg et al. (2018), p. 26. A higher cost for mimicking, those authors explain, results in a separating equilibrium with a highly informative digital footprint. They illustrate this with the example of Pentaquark, who rejects loans from applicants who "write a lot about their souls on Facebook, as these persons are usually too concerned about what will happen in thirty years, but not the fine print of today's life".
[318] Berg et al. (2018), pp. 26–27.
[319] CFPB (2017), p. 17; see on further concerns, such as "gaming the system"; Burk (2021), pp. 1187 et seq.; Citron/Pasquale (2014), pp. 29 et seq.; Langenbucher (2020).
[320] Below F.III.2; FTC (2022) p. 51275.
[321] Art. 13 para. (2) lit. f GDPR.
[322] Art. 15 GDPR.
[323] Art. 18 para. 6.

Another concern are proprietary trade secrets. Courts have not required lenders or scoring agencies to disclose the model they use to evaluate creditworthiness.[324] Scoring agencies in the US have long disclosed various attributes they consider. This gives consumers the chance to work on the relevant variables to better their score.[325] With algorithms and big data, scoring agencies might have less incentives to do so. A scoring agency that found attributes which are particularly predictive of credit default risk, for instance having a dating or a finance app installed, has little interest in disclosing that. The fear is that as soon as potential borrowers become aware, they will react and delete (or install) the relevant app. For the scoring agency, this means losing a variable which was easy to establish and had good predictive force.[326] Additionally, not all models allow for reverse engineering to find out which variables were given most weight.

2. Controlling Optimization Goals

One element of proprietary trade secrets are optimization goals. They define what the algorithm is looking for.[327] In theory, lenders benchmark their offer against market prices for comparable loans and look for the most competitive price they can refinance. In this case, their optimization goal is credit default risk to build groups of borrowers. Under this assumption, one would expect comparable terms and conditions across lenders when faced with similarly situated applicants. Bank regulators as well as courts are likely to encourage, rather than limit the use of AI underwriting models to that end. This seems particularly appealing if algorithms contribute to more granular risk assessment, better risk-adjusted pricing, and in this way contribute to sound and stable financial markets. It remains to be seen to what extent moral concerns about using sensitive data, such as those implicit in the EU Consumer Credit Directive/2021,[328] will enter into the equation.

In practice, things are less straightforward. Empirical studies have highlighted inequality in output even for similarly situated persons.[329] Borrowers vary in their access to information, their financial literacy, and the urgency of their need for credit. As Fuster and his co-authors have speculated,[330] this could explain unequal outcome across groups which does not track variation in credit default risk.[331] On the side of the lender, anti-competitive practices could keep interest rates at a higher level than economic theory predicts. Regulatory arbitrage with local laws on interest rate caps might facilitate predatory lending.[332] Against that background, it is worrying to understand that an AI

[324] Langenbucher (2020), p. 542.
[325] See above D.
[326] Langenbucher (2020), pp. 542 et seq.
[327] O'Neill (2016), p. 21 ("definition of a success").
[328] See above F.II.1.
[329] See above B.
[330] See above E.I.
[331] Hurlin et al. (2021) offer a methodology to distinguish whether a lender discriminates only for creditworthiness.
[332] In the US, a federal bank can originate loans across all states with the highest interest rate permissible under the law of the state where it is headquartered ("exportation doctrine"). Utah has been prominent for a state without a ceiling on interest rates. Under EU law, Art. 31 of the EU Consumer Credit Directive/2021includes the obligation of Member States to set an interest rate cap (without specifying a ceiling). For Germany, see sec. 138 *Bürgerliches Gesetzbuch* (German Civil Code) and longstanding jurisprudence which voids a contract where interest rates are more than 90 % higher than the market price in the relevant segment.

model's optimization goal can not only assist in finding invisible primes.[333] It can also be helpful in identifying applicants which are likely to sign a loan above market price, for instance because they do not have the time or skill for comparison shopping.[334] Vulnerable borrowers will often not be aware that the access to their data they provide hurts, rather than helps them.

Lenders will not usually be open to sharing their optimization goals. Additionally, fitting prices to a specific audience is not prohibited *per se* but an element of free contracting. Standard exceptions have to do with antitrust law or with unfair business practices.[335] The EU Consumer Credit Directive/2021 explicitly encourages lenders to "personalize the price of their offers for specific consumers or specific categories of consumers based on automated decision-making and profiling of consumer behavior allowing them to assess the consumer's purchasing power".[336] The EU Digital Markets Act restricts some forms of targeted advertising, but only for gatekeepers. Personalized pricing can sometimes be an economically efficient price-seeking mechanism and increase competition.[337] At the same time, whenever profiling along these lines targets vulnerable applicants to extend predatory loans, it will often be an inefficient rent-seeking activity.[338] Additionally, if consumers and firms react by trying to change their online behavior,[339] welfare losses are likely.[340]

Against this background, the openness of the current legal framework towards personalized pricing calls for a normative double-check of the type mentioned above.[341] In this paper, I do not engage with the discussion on personalized pricing but offer only brief remarks to indicate the need for further work.[342] Personalized prices are a form of first-degree price discrimination if they target individual consumers based on their preferences and reservation values.[343] They are impermissible, if they imply a discriminatory business strategy.[344] Under the FHA and the ECOA, the US DoJ has claimed that this also true for practices such as "reverse redlining". This argument holds even though anti-discrimination laws typically respond to denying a loan, not to granting an expensive one.[345]

While the balance regulators and legislators must strike in that space is not a novel one, AI compounds existing concerns in two ways. The first is scale. Big data ML models make it less costly to find various vulnerable groups, hence have the potential to amplify the problem. Additionally, many

---

[333] See above B.
[334] Aggarwal (2021), p. 50.
[335] Ernst (2017), p. 1034 on Art. 22 GDPR *ibid*, p. 1034-1035.
[336] Recital (40); on profiling under the GDPR see Kaminski (2019), p. 1551; Langenbucher (2020), pp. 538, 540.
[337] Eidenmüller/Wagner (2021), pp. 53; Ernst (2017), p. 1034.
[338] Aggarwal (2021), p. 50; Eidenmüller/Wagner (2021), pp. 50-54.
[339] See F.III.1.
[340] See Eidenmüller/Wagner (2021), p. 53 for a preliminary summary: "If anything can be said with reasonable certainty (…) it is that, in the aggregate, first-degree price discrimination benefits firms and harms consumers. The overall net effects are unclear".
[341] See E.II.2.
[342] See in the context of AI Eidenmüller/Wagner (2021), pp. 47-71.
[343] Eidenmüller/Wagner (2021), pp. 51.
[344] Above E.I.
[345] Available at https://www.justice.gov/crt/housing-and-civil-enforcement-cases-documents-231 (last access 27 October 2022).

targeted offers of this type will escape liability under anti-discrimination laws altogether.[346] Take, for instance, lenders which target students, recent immigrants, or refugees, having found that they are likely to accept a higher mark-up on prices than the average borrower. These lenders do not qualify under anti-discrimination laws unless the loan portfolio is skewed towards protected characteristics. Even if regulators can require lenders to disclose proprietary business strategies and optimization goals, the algorithm might optimize high-level goals, such as meeting a profit goal. Regulators performing a model check will then need deep access, for instance to run their own AI, charged with understanding what drives the specific lender's model.

When dealing with these intricacies, one regulatory strategy looks to transparency.[347] Informing borrowers that an underwriting decision takes personal attributes of the applicant into account will not come as a surprise to most applicants.[348] However, this is different if applicants understand that prices are personalized even for comparably situated borrowers. A regulatory requirement to disclose this to applicants might set incentives for more comparison shopping, thereby enhancing competition between lenders. The EU Consumer Credit Directive/2021 provides an illustration, requiring lenders to inform consumers if profiling is used and to provide for human intervention. Additionally, lenders must provide a clear explanation of the individual assessment and the possibility to contest the creditworthiness assessment if the borrower requests it.[349]

Another regulatory avenue concerns responsible lending principles.[350] These have the individual borrower's financial situation in mind; hence, they go beyond the definition of protected groups under anti-discrimination law. Additionally, their policy goals aim at ensuring the stability of the financial system. Against this background, regulators would arguably be in a strong position to request access to optimization goals on loans and interest rates with both, over-indebtedness of individual borrowers and macro-stability of the financial system in mind. The EU Consumer Credit Directive/2021 illustrates this by requiring that a credit default risk check "shall be done in the interest of the consumer, to prevent irresponsible lending practices and overindebtedness".[351]

IV. The Cost of Equal Access

The considerable predictive power of AI models has surfaced throughout this paper. Increasingly, they provide a more granular picture of individual applicants than traditional models. The hope is to exploit large data pools by machines, rather than cognitively limited humans. Under that assumption, each borrower would be presented with a custom-made offer of credit, as it were. Hopes such as these have led some scholars to suggest that in the future, regulating inequality in access to credit will boil

[346] Ernst (2017), p. 1034.
[347] Ernst (2017), pp. 1034-1035.
[348] This can be different in mass retail credit situations.
[349] Art. 12, 13, 18 para. 6.
[350] See the EU Consumer Credit Directive/202 Art. 18, recital (45), (46) and the current Consumer Credit Directive 2008 recital (26).
[351] Art. 18 para. (1) EU Consumer Credit Directive/2021; other examples at Eidenmüller/Wagner (2021), pp. 54-55.

down to a balance between accuracy and fairness. As Sunstein claims, "use of algorithms will reveal, with great clarity, the need to make tradeoffs between the value of racial (or other) equality and other important values".[352]

This is a tempting picture. But we are not yet there. Data can be incorrect at the outset, a concern particularly relevant for social media data.[353] Data can be correct when collected but reflect an outdated picture.[354] Data points can carry a meaningful message for many, but not all persons.[355] Algorithms which use such data do not produce an accurate picture of all applicants. What is more, models optimize a definition of success.[356] In that pursuit they make their own tradeoffs long before they deliver their assessment of an individual applicant to the user of the algorithm. This is not to suggest that human credit officers are doing a better job judging borrowers, but to make a simple point: algorithms do not deliver a quick fix to calculating the costs of equal access to credit. As is true for most normative decisions regulators and legislator make, uncertainty about facts and about future developments is part of the challenge.

Against this background, the paper suggests a further double-check of the current normative framework.[357] The focal points of US and EU laws in that space vary, due to culture and history. The US has, under the ECOA and the FHA, a statutory tradition in ensuring equal access to credit via federal law. By contrast, regulating predatory lending is done by states and varies significantly. The EU has under the GDPR emphasized privacy, but not access to credit. Predatory lending practices vary across EU Member States. The EU Consumer Credit Directive/2021 takes a first step towards harmonization of caps on predatory lending practices and on discrimination in credit underwriting.[358] The GDPR and the FCRA both highlight transparency for consumers.[359] After the financial crisis of 2008, regulators have globally pushed for responsible lending practices.[360]

The lively debate among economists, legal and information systems scholars has focused on different strategies to adjust the outcome of algorithmic decision-making. One way to cope is tuned to algorithmic fairness. Hurlin et al. provide an illustration for credit underwriting. They develop a formal definition for fairness based on statistical parity, equal odds, predictive parity, and overall accuracy. Individual lenders can benchmark their own models against a fairness model along those

[352] Sunstein (2019), p. 504.
[353] See F.II.2.
[354] See D.
[355] See D.III.
[356] See F.III.2.
[357] See above F.III.2. for the first double-check the paper advocated as to controlling business strategies of lenders.
[358] Art. 31 requires Member States to introduce caps on interest rates an total cost but leaves it to the Member States to decide on the amount of the ceiling.
[359] Langenbucher (2020).
[360] See the EU EBA guidelines on loan originating and monitoring, available at: https://www.eba.europa.eu/regulation-and-policy/credit-risk/guidelines-on-loan-origination-and-monitoring (last access 27 October 2022); EU Council Action Plan on tackling non-performing loans, available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2375 (last access 27 October 2022).

lines.[361] Gillis proposes instead to go for a pure outcome control.[362] Lenders decide on the pricing model they wish to use. They submit it to the regulator to test. The regulator has prepared a model portfolio of borrowers and runs the lender's algorithm to see if it triggers disparities across protected groups. If it does, the regulator decides whether these are acceptable. Rolling this proposal out more broadly, any lender using underwriting algorithms would be required to undergo regulatory scrutiny by running his proprietary AI model on the portfolio the regulator has established. If the outcome was in the bandwidth set by the regulator, the lender would be free to use his model. To enhance transparency for consumers, the regulator might even offer "test runs" for potential borrowers, to learn how a change in input variable affects their current score.[363] Skeptics of output controls along those lines have raised concerns as to gaming the system. The more information coders receive about the model benchmark, the more likely the algorithm will learn to produce the expected result without necessarily achieving the type of equality in access which the regulator had in mind.[364] Yet others push output control even further, calling for "algorithmic affirmative action".[365]

It is beyond the scope of this paper, focused on the limits of received anti-discrimination doctrine, to engage with this debate. In follow-up work I hope to show advantages of normative interventions targeted to the specific area of AI usage. Credit scoring differs from other forms of algorithmic decision-making, hence, calls for its own, custom-made answers. To illustrate, I have pointed to problems of detecting false negatives in underwriting decisions. While algorithms are in a good position to cope with these in, for example, medicine, this is different in credit underwriting.[366] Another argument why targeted, "siloed" interventions might be attractive looks to the interests and incentives of actors on credit markets. Leveraging these can help to find more finely tuned answers than static benchmark models or output controls. Regulating access to consumer credit is for many a means to offer equal chances and opportunities. Taking out a loan can help to cover needs which the state does not, ranging from health care over unemployment aid to student tuition. Providing for equal opportunities is an especially acute concern in those areas. At the same time, relaxing standards for credit default risk not only hurts shareholders and stakeholders of the lender. It also risks instability in the financial system if lower standards lead to credit bubbles as seen preceding the financial crisis of 2008. One reaction is the approach of the EU Consumer Credit Directive/2021: Even with a negative creditworthiness assessment, credit can be made available, but only to fund exceptional healthcare expenses, student loans or loans for consumers with disabilities.[367] Another approach could enlist government-sponsored entities such as US *Fannie Mae* and *Freddie Mac* which could guarantee consumer loans in specific cases. Striking a balance between the various competing interests will look very different across countries, cultures, historical heritage, and institutional design of regulators. Solutions will often be preliminary and require public debate. The challenge for credit scoring will be to find the right balance between the competing interests of private actors, stability

[361] Hurlin et al. (2021), pp. 5, 11, 16.
[362] Gillis (2022), pp. 67 et seq.
[363] See for a suggestion along these lines Citron/Pasquale (2014), pp. 28 et seq.
[364] For a discussion of "gaming the system" see Citron/Pasquale (2014), pp. 29 et seq.
[365] Chander (2017), p. 1039; for a critique see Mayson (2019), pp. 2267 et seq.
[366] See above D.III.
[367] Recital (47). Of course, this raises the concern of predatory lending, see F.III.2., again making the control of business strategies of lenders an important concern.

concerns of the financial system and the interest in providing a safe environment for responsible innovation.

G. SUMMARY

The potential of big data and AI credit underwriting models to lower search costs for lenders marks the introduction to this paper. Going beyond standard metrics, algorithms can help to identify invisible primes, and there is considerable empirical evidence on the achievements of Fintech companies which are active in this market. Under ideal market conditions, economists would expect that variation in access to credit can be explained by disparity in credit risk. Under this assumption, AI models and big data provide a useful tool to make more granular predictions than traditional metrics. However, empirical analysis on US markets suggests that Fintech algorithms tend to be more advantageous for some minorities, such as white Hispanics and Asians, than for others, such as the non-white Hispanic and Black population. Among the hypotheses which try to explain these findings, an especially worrying one points to strategic pricing. Algorithms can identify groups in more urgent need for credit, hence, more likely to accept less favorable conditions than similarly situated groups. The paper gave an overview on empirical papers in that area. Biases of training data and of algorithms are further reasons for the disparate output we find. The paper summarized findings of computer scientists in that space.

Against that backdrop, the paper tried to show that received anti-discrimination doctrine is ill-suited to deal with algorithmic discrimination. One of the concerns which have sparked the current debate are proxies. A proxy variable is neutral on its face but correlates with a protected attribute. While this is not *per se* a new phenomenon, AI has the potential to strongly amplify the problem. Big data provides a universe of possible proxies and self-learning models identify correlations with ease and precision. This leads to frictions with the law's traditional understanding of human decision making, which is rooted in distinct building blocks along a chain of causation. Many of these building blocks are harmless, lawful motivations. Some concern protected attributes and are considered outright unlawful. A few are facially neutral but suspicious proxies, standing in for a protected attribute. Implicit is the understanding that there is a limited number of proxies available for human decision-makers. I have submitted that, in the face not only of a massive increase of potential proxies, but also of bundles of variables which, taken together, accurately predict a protected attribute, anti-discrimination law's received concept of causation will lose significance.

Disparate treatment doctrine handles decisions which are made *because of* protected attributes. While this sounds straightforward at first glance, this paper has looked to what was traditionally called discrimination by proxy. These are cases where the building blocks or practices for a decision masked a protected attribute which was the real motive. Against this background, I have asked where masking ends and disparate impact starts. While the question is not necessarily a novel one, AI models with their multitude of variables compound the underlying difficulty to establish what can be considered the motive for a decision. I have joined the prevailing opinion in not requiring intent when a protected attribute explicitly forms part of the building blocks of a decision. By contrast, intent is a necessary element in traditional discrimination by proxy cases, where the discriminator deliberately hides behind a seemingly neutral attribute.

Against that background, this paper has understood disparate impact as the main battleground for algorithmic discrimination. Received doctrine links the disparate outcome for a protected group to facially neutral attributes which cause the discriminatory decision. Causation can be established if removing the variable and leaving all other factors constant leads to a non-discriminatory outcome.

The paper's main argument when dealing with disparate impact centered on this element of causation. Eliminating the offensive variable, so it submitted, will change the outcome only for limited-input models. For sophisticated algorithms, due to redundant encoding and to the flexibility of multivariate regressions, the model will use stand-in proxies to arrive at the same prediction. This makes the received but-for causation test unfit to cope with many algorithmic models.

The paper moved on to explore whether a solution lies in understanding the entire AI model as the building block which causes the decision. The problem with this approach was that it neither worked for fully automated models nor for human credit officers which relied primarily on the algorithmic recommendation. For these situations, removing the entire model means there are no other factors left which cause the decision. Put differently: the logic of a causation test requires removing the building block and understanding what would have happened without it. Removing the entire AI model makes it impossible to go through with this test, if it is the algorithm (not a human credit officer) which caused the underwriting decision.

It was interesting to see that the US CFPB had approached the problem in a different manner. Instead of investigating the distinct variables fed into the algorithm to establish causation, the Bureau came up with its own hypothetical counterfactual. It compared the absolute number of minority borrowers the algorithm recommended to the absolute number a hypothetical FICO score model would have recommended.

The paper understood this approach as an illustration of the need to move beyond anti-discrimination law when dealing with algorithmic scoring and credit underwriting. The causation element is meaningful in a world of human decision-making where people hide true intentions or need incentives to look for non-discriminatory strategies. It is less helpful for algorithmic decision-making. Against this background, the paper outlined preliminary contours of a regulatory design of fair lending in the age of AI. Details of this design will be the topic of later papers.

One element of the regulatory design I proposed is quality control. This includes technical and governance controls, both as to data and model. Flaws can hurt the borrower, if he is rejected or overpays, and the lender, if he leaves money on the table by refusing a loan which would have been attractive. The EU AI Act illustrates avenues towards quality control of this type.

Another step towards a regulatory design looked towards credit reporting and financial privacy. Information gathered from sources such as social media networks are prone to mistakes and misunderstandings, data points might include sensitive information. Current laws provide consumers with various degrees of rights to access their data, ask for rectification and erasure. Some, such as the US FCRA, focus on the underwriting context. Others, such as the GDPR, work with omnibus rules.

The paper submitted that both sets of laws need adjustment to the age of AI credit underwriting, especially with an eye on efficient enforcement.

The paper moved on to explore legislative decisions from a pre-AI period in need of a normative double-check. One such area has to do with striking the right balance between a borrower's interest in transparency of algorithmic decision-making and the lender's proprietary trade secrets. Algorithmic underwriting decisions are often opaque for borrowers, sometimes even for the lender. This makes it difficult to establish trust in AI models, if applicants feel they lose autonomy given that an inscrutable algorithm assigns them scores which they cannot fully apprehend. One element of proprietary trade secrets are the lender's optimization goals. These define what the algorithm looks for. An evaluation of credit default risk is a somewhat natural optimization goal. However, an algorithm might instead search for vulnerable applicants, prone to accepting predatory loans. In that space, the paper hinted at the need for a normative double-check on current laws. They were understood as too liberal in accepting personalized pricing to the detriment of vulnerable groups. The paper suggested to include responsible lending principles in this endeavor, aiming towards regulatory control of optimization goals.

The paper concluded with another normative double-check. Believers in the objectivity of algorithmic underwriting models tend to claim that these present prices as they "should be". Any deviation from these prices produces costs and AI models spell them out. I stressed that this is a tempting, but as of today not a realistic picture. Incorrect and biased data produce flawed models which do not present an accurate picture of reality. For now, the challenge is not only to find meta-reasons of distributive justice to reallocate these hypothetical costs. Rather, the paper suggested to explore targeted, finely tuned interventions. These could include credit guarantees by government-sponsored entities such as Freddie Mac or Fannie Mae in the US or EU plans to enable credit for applicants in a situation of hardship.

References

Adams-Prassl, Jeremias/Binns, Reuben/Kelly-Lyth, Aislinn (2022): Directly Discriminatory Algorithms, Modern Law Review, Vol. 85, preprint available at https://onlinelibrary.wiley.com/doi/full/10.1111/1468-2230.12759 (last access 27 October 2022).

Agarwal, Sumit/Shashwat, Alok/Ghosh, Pulak/Gupta, Sudip (2021): Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech, Indian School of Business, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3507827 (last access 27 October 2022).

Aggarwal, Nikita (2021): The Norms of Algorithmic Credit Scoring, The Cambridge Law Journal, Vol. 80, pp. 42–73.

Al-Zuabi, Ibrahim Mousa/Jafar, Assef/Aljoumaa, Kadan (2019): Predicting customer's gender and age depending on mobile phone data, Journal of Big Data, available at https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0180-9 (last access 27 October 2022).

American Bankers Association (ABA) (2022): Petition for rulemaking defining larger participants of the aggregation services, available at https://www.aba.com/-/media/documents/letters-to-congress-and-regulators/petition-to-cfpb-for-larger-participant-rulemaking-080222.pdf?rev=c0674598829b40808a179b1f4942b591 (last access 27 October 2022).

Arrow, Kenneth J. (1971): Some models of Racial Discrimination in the Labor Market, available at https://apps.dtic.mil/sti/pdfs/AD0735068.pdf (last access 27 October 2022).

Awrey, Dan/Macey, Joshua (2022): The Promise and Perils of Open Finance, ECGI working paper No 632/2022, available at https://ecgi.global/sites/default/files/working_papers/documents/maceyawreyfinal.pdf (last access 27 October 2022).

Balyuk, Tetyana (2021): FinTech Lending and Bank Credit Access for Consumers, Rotman School of Management Working Paper No. 2802220, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2802220 (last access 27 October 2022).

Barocas, Solon/Selbst, Andrew D. (2016): Big Data's Disparate Impact, California Law Review, Vol. 104, pp. 671–732.

Bartlett, Robert/Morse, Adair/Stanton, Richard/Wallace, Nancy (2022): Consumer-Lending Discrimination in the FinTech Era, Journal of Financial Economics Vol. 143, pp. 30-56.

Becker, Gary S. (1957): The Economics of Discrimination, Chicago University Press.

Becker, Gary S. (1993): Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education, Chicago University Press, available at https://doi.org/10.7208/chicago/9780226041223.001.0001 (last access 27 October 2022).

Benjamin, Ruha (2019): Race after Technology, Polity Press.

Berg, Tobias/Burg, Valentin/Gombović, Ana/Puri, Manju (2018): On the Rise of FinTechs – Credit Scoring using Digital Footprints, NBER Working Papers 24551, National Bureau of Economic

Research, Inc., available at https://www.nber.org/system/files/working_papers/w24551/w24551.pdf (last access 27 October 2022).

Berman, Mitchell N./Krishnamurthi, Guha (2021): Bostock was Bogus: Textualism, Pluralism, and Title VII, Notre Dame Law Review, Vol. 97, pp. 67–126.

Blattner, Laura/Nelson, Scott (2021): How Costly is Noise? Data and Disparities in Consumer Credit, available at https://www.researchgate.net/publication/351656623_How_Costly_is_Noise_Data_and_Disparities_in_Consumer_Credit (last access 27 October 2022).

Bordalo, Pedro/Coffman, Katherine/Gennaioli, Nicola/Shleifer, Andrei (2016): Stereotypes, Quarterly Journal of Economics, Vol. 131, pp. 1753–1794.

Brito, Dagobert/Hartley, Peter (1995): Consumer Rationality and Credit Cards, Journal of Political Economy, Vol. 103, pp. 400–433.

Bureau of Consumer Financial Protection (CFPB) (2017): Request for Information, Docket-nr. CFPB-2017-0005 (2017), available at https://files.consumerfinance.gov/f/documents/20170214_cfpb_Alt-Data-RFI.pdf (last access 27 October 2022).

Bureau of Consumer Financial Protection (CFPB) (2019): An update on credit access and the Bureau's first No-Action Letter, available at https://www.consumerfinance.gov/about-us/blog/update-credit-access-and-no-action-letter/ (last access 27 October 2022).

Bureau of Consumer Financial Protection (CFPB) (2020a), available at https://files.consumerfinance.gov/f/documents/cfpb_upstart-network-inc_no-action-letter_2020-11.pdf (last access 27 October 2022)

Bureau of Consumer Financial Protection (CFPB) (2020b): Consumer Access to Financial Records, 85 FR 71003, pp. 71003–71011.

Burk, Dan L. (2021): Algorithmic Legal Metrics, Notre Dame Law Review, Vol. 96, pp. 1147–1203.

Burrell, Jenna/Fourcade, Marion (2021): The Society of Algorithms, Annual Review of Sociology, Vol. 47, pp. 213–237.

Campbell, Colin/Smith, Dale (2022): Distinguishing Between Direct and Indirect Discrimination, Modern Law Review, Vol. 85, preprint available at https://onlinelibrary.wiley.com/doi/full/10.1111/1468-2230.12760 (last access 27 October 2022).

Chander, Anupam (2017): The Racist Algorithm?, Michigan Law Review, Vol. 115, pp. 1023–1045.

Citron, Danielle K./Pasquale, Frank (2014): The Scored Society: Due Process for Automated Predictions, Washington Law Review, Vol. 89, pp. 1–32.

Dembroff, Robin/Kohler-Hausmann, Issa (2022): Supreme Confusion About Causality at the Supreme Court, City University of New York Law Review, Vol. 25, pp. 57–92.

DeYoung, Robert/Phillips, Ronnie J. (2006): Strategic Pricing of Payday Loans: Evidence from Colorado, 2000–2005, available at https://www.indstate.edu/business/sites/business.indstate.edu/files/Docs/2006-WP-05_Young-Phillips.pdf (last access 27 October 2022).

Di Maggio, Marco/Ratnadiwakara, Dimuthu/Carmichael, Don (2021): Invisible Primes: Fintech Lending with Alternative Data, Harvard Business School Working Paper No. 22-024, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3937438 (last access 27 October 2022).

Dobbie, Will/Liberman, Andres/Paravisini, Daniel/Pathania, Vikram (2019): Measuring Bias in Consumer Lending, Harvard Kennedy School Working Paper No. RWP19-029, available at https://www.hks.harvard.edu/publications/measuring-bias-consumer-lending (last access 27 October 2022).

Dzida, Boris/Groh, Naemi (2018): Diskriminierung nach dem AGG beim Einsatz von Algorithmen im Bewerbungsverfahren, NJW, pp. 1917–1922.

Eidenmüller, Horst/Wagner, Gerhard (2021): Law by Algorithm, Mohr Siebeck.

European Data Protection Board/European Data Protection Supervisor (EDPB/EDPS) (2021): Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act), 18 June 2021, avalaible at https://edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf (last access 27 October 2022).

Eidelson, Benjamin (2022): Dimensional Disparate Treatment, Southern California Law Review (forthcoming), Vol. 95, pp. 785–855.

Ernst, Christian (2017): Algorithmische Entscheidungsfindung und personenbezogene Daten, Juristenzeitung, pp. 1026-1036.

Federal Trade Commission (FTC) (2011): 40 years of experience with the FCRA, 2011, available at https://www.ftc.gov/sites/default/files/documents/reports/40-years-experience-fair-credit-reporting-act-ftc-staff-report-summary-interpretations/110720fcrareport.pdf (last access 27 October 2022).

Federal Trade Commission (FTC) (2022): Advance notice of proposed rulemaking, Trade Regulation Rule on Commercial Surveillance and Data Security, 87 FR 51273, pp. 51273–51299.

Financial Stability Board (FSB) (2022), Fintech and Market Structure in the COVID-19 Pandemic, 21 March 2022, available at https://www.fsb.org/wp-content/uploads/P210322.pdf (last access 27 October 2022).

Fisher, Linda E. (2009): Target Marketing of Subprime Loans: Racialized Consumer Fraud & Reverse Redlining, Journal of Law and Policy, Vol. 18, pp. 121–155.

Fuster, Andreas/Goldsmith-Pinkham, Paul/Ramadorai, Tarun/Walther, Ansgar (2022): Predictably Unequal? The Effects of Machine Learning on Credit Markets, Journal of Finance, Vol. 77, pp. 5–47.

Gillis, Talia (2022): The Input Fallacy, Minnesota Law Review (forthcoming), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3571266 (last access 27 October 2022).

Graham, Jennifer (2021): Risk of discrimination in AI systems, Evaluating the effectiveness of current legal safeguards in tackling algorithmic discrimination, in: Lui/Ryder (ed.), FinTech, Artificial Intelligence and the Law, pp. 211–229.

Guseva, Alya/Rona-Tas, Akos (2001): Uncertainty, Risk, and Trust: Russian and American Credit Card Markets Compared, American Sociological Review, Vol. 66, pp. 623–646.

Hacker, Philipp (2018): Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law, Common Market Law Review, Vol. 55, pp. 1143–1185.

Hellman, Deborah (2020): Measuring Algorithmic Fairness, Virginia Law Review, Vol. 106, pp. 811–866.

Hurlin, Christophe/Pérignon, Christophe/Saurin, Sébastien (2021): The Fairness of Credit Scoring Models, HEC Paris Research Paper No. FIN-2021-1411, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3785882 (last access 27 October 2022).

Kaminski, Margot (2019): Binary Governance: Lessons from the GDPR's approach to Algorithmic Accountability, Southern California Law Review, Vol. 92, pp. 1529–1616.

Kim, Pauline (2022): AI and Inequality, in: Johnson, Kristin/Reyes, Carla (ed.), The Cambridge Handbook on Artificial Intelligence and the Law (forthcoming), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3938578.

Kim, Pauline/Hanson, Erika (2016): People Analytics and the Regulation of Information under the Fair Credit Reporting Act, St Louis University Law Journal, Vol. 61, pp. 17–34.

Kissinger, Henry/Schmidt, Eric/Huttenlocher, Daniel (2021): The Age of AI: And Our Human Future, Little Brown and Company.

Kiviat, Barbara (2019a): The Art of deciding with data: evidence from how employers translate credit reports into hiring decisions, Socio-Economic Review, Vol. 17, pp. 283–309.

Kiviat, Barbara (2019b): The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores, Socio-Economic Review, Vol. 84, pp. 1134–1158.

Koppelman, Andrew (2022): Bostock and Textualism: A Response to Berman and Krishnamurthi, Notre Dame Law Review Reflection (forthcoming), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3904246 (last access 27 October 2022).

Krishnamurti, Guha/Salib, Peter (2020): Bostock and Conceptual Causation, Yale Journal of Regulation, Notice & Comment, available at https://www.yalejreg.com/nc/bostock-and-conceptual-causation-by-guha-krishnamurthi-peter-salib/ (last access 27 October 2022).

Langenbucher, Katja (2020): Responsible A.I. credit scoring – a legal framework, European Business Law Review, Vol. 31, pp. 527–572.

Langenbucher, Katja (2022): AI credit scoring and evaluation of creditworthiness – a test case for the EU proposal for an AI Act, in ECB, Continuity and change – how the challenges of today prepare the ground for tomorrow, ECB Legal Conference 2021, pp. 362–386, available at

https://www.ecb.europa.eu/pub/pdf/other/ecb.ecblegalconferenceproceedings202204~c2e5739756.en.pdf (last access 27 October 2022).

Langenbucher, Katja/Corcoran, Patrick (2022): Responsible AI Credit Scoring – A Lesson from Upstart.com, European Company and Financial Law Review, Vol. 5, pp. 141–179.

Lauer, Josh (2017): Creditworthy, Columbia University Press.

Ledford, Heidi (2019): Millions of black people affected by racial bias in health-care algorithms, Nature Portfolio, available at https://www.nature.com/articles/d41586-019-03228-6 (last access 27 October 2022).

Lucas, Robert (1976): Econometric policy evaluation: A critique, Carnegie-Rochester Conference Series on Public Policy, Vol. 1, pp. 19–46.

Mayson, Sandra (2019): Bias In, Bias Out, The Yale Law Journal, Vol. 128, pp. 2218–2300.

Megginson, W. L., Lopez, D., & Malik, A. I. (2021). The rise of state-owned investors: sovereign wealth funds and public pension funds. Annual Review of Financial Economics, 13, 247-270.

National Consumer Law Clinic (NCLC) (2020): Written Statement for CFPB's Symposium on Consumer Access to Financial Records, available at https://www.nclc.org/resources/nclc-statement-for-cfpb-1033-symposium/ (last access 27 October 2022).

Obermeyer, Ziad/Powers, Brian/Vogeli, Christine/Mullainathan, Sendhil (2019): Dissecting racial bias in an algorithm used to manage the health of populations, Science, Vol. 336, pp. 447–453.

O'Keefe, Patrick (2016): Qualified Mortgages & Government Reverse Redlining: How the CFPB's Qualified Mortgage Regulations Will Handicap the Availability of Credit to Minority Borrowers, Fordham Journal of Corporate & Financial Law, Vol. 21, pp. 413–447.

O'Neil, Cathy (2016): Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Penguin Books.

Parlour, Christine/Rajan, Uday (2001): Competition in Loan Contracts, The American Economic Review, Vol. 91, pp. 1311–1328.

Phelps, Edmund (1972): The Statistical Theory of Racism and Sexism, The American Economic Review, Vol. 62, pp. 659–661.

Prince, Anya E.R./Schwarcz, Daniel (2020): Proxy Discrimination in the Age of Artificial Intelligence and Big Data, Iowa Law Review, Vol. 105, pp. 1257–1318.

Sacksofsky, Ute (2017): Was heißt: Ungleichbehandlung „wegen"?, in: Kempny/Reimer (ed.), Gleichheitssatzdogmatik heute, pp. 63–90.

Stiglitz, Joseph/Weiss, Andrew (1981): Credit Rationing in Markets with Imperfect Information, The American Economic Review, Vol. 71, pp. 393–410.

Student Borrower Protection Center (2020), Educational Redlining, February 2020, available at https://protectborrowers.org/wp-content/uploads/2020/02/Education-Redlining-Report.pdf (last access 27 October 2022).

Sunstein, Cass (2019): Algorithms, Correcting Biases, Social Research: An International Quarterl Vol. 86, pp. 499–511.

Wachter, Sandra (2022): The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law, Tulane Law Review (forthcoming), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4099100 (last access 27 October 2022).

Wachter, Sandra/Mittelstadt, Brent/Russell, Chris (2021): Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI, Computer Law & Security Review, Vol. 41, available at: https://www.sciencedirect.com/science/article/abs/pii/S0267364921000406 (last access 27 October 2022).

Zuiderveen Borgesius, Frederik (2020): Price Discrimination, Algorithmic Decision-Making, and European Non-Discrimination Law, European Business Law Review, Vol. 31, pp. 401–422.

about ECGI

The European Corporate Governance Institute has been established to improve *corporate governance through fostering independent scientific research and related activities.*

The ECGI will produce and disseminate high quality research while remaining close to the concerns and interests of corporate, financial and public policy makers. It will draw on the expertise of scholars from numerous countries and bring together a critical mass of expertise and interest to bear on this important subject.

The views expressed in this working paper are those of the authors, not those of the ECGI or its members.

www.ecgi.global

ECGI Working Paper Series in Law

https://ecgi.global/content/working-papers

Electronic Access to the Working Paper Series

The full set of ECGI working papers can be accessed through the Institute's Web-site (https://ecgi.global/content/working-papers) or SSRN:

Finance Paper Series	http://www.ssrn.com/link/ECGI-Fin.html
Law Paper Series	http://www.ssrn.com/link/ECGI-Law.html