Algorithmic Harm in Consumer Markets

Oren Bar-Gill,[*] Cass R. Sunstein[**] and Inbal Talgam-Cohen[***]

Abstract

Machine learning algorithms are increasingly able to predict what goods and services particular people will buy, and at what price. It is possible to imagine a situation in which relatively uniform, or coarsely set, prices and product characteristics are replaced by far more in the way of individualization. Companies might, for example, offer people shirts and shoes that are particularly suited to their situations, that fit with their particular tastes, and that have prices that fit their personal valuations. In many cases, the use of algorithms promises to increase efficiency and to promote social welfare; it might also promote fair distribution. But when consumers suffer from an absence of information or from behavioral biases, algorithms can cause serious harm. Companies might, for example, exploit such biases in order to lead people to purchase products that have little or no value for them or to pay too much for products that do have value for them. Algorithmic harm, understood as the exploitation of an absence of information or of behavioral biases, can disproportionately affect identifiable groups, including women and people of color. Since algorithms exacerbate the harm caused to imperfectly informed and imperfectly rational consumers, their increasing use provides fresh support for existing efforts to reduce information and rationality deficits, especially through optimally designed disclosure mandates. In addition, there is a more particular need for algorithm-centered policy responses. Specifically, algorithmic transparency—transparency about the nature, uses, and consequences of algorithms—is both crucial and challenging; novel methods designed to open the algorithmic "black box" and "interpret" the algorithm's decision-making process should play a key role. And, in appropriate cases, regulators should police the design and implementation of algorithms, with a particular emphasis on exploitation of an absence of information or of behavioral biases.

[*] William J. Friedman and Alicia Townsend Friedman Professor of Law and Economics, Harvard Law School.

[**] Robert Walmsley University Professor, Harvard Law School.

For helpful comments and conversations, we thank Omri Ben-Shahar, Ben Eidelson, Talia Gillis, Shafi Goldwasser, Assaf Hamdani, Sharon Hannes, Howell Jackson, Louis Kaplow, Emiliano Katan, Tamar Kricheli-Katz, Haggai Porat, Lucia Reisch, Ricky Revesz, Sarath Sanga, Alan Schwartz, Steve Shavell, Yonadav Shavit, Holger Spamann, Rory Van Loo, and workshop and conference participants at Harvard, Tel-Aviv University, […] and at the 2022 Annual Meeting of the American Law and Economics Association. Rachel Neuburger and Davy Perlman provided excellent research assistance.

[***] Assistant Professor, Technion – Israel Institute of Technology, The Henry and Marilyn Taub Faculty of Computer Science.

TABLE OF CONTENTS

I. Introduction

Sellers and service providers are increasingly using machine learning algorithms.[1] Many uses should greatly benefit consumers. Suppose that algorithms can predict what goods and services people will buy and at what price. If algorithms give people information about beneficial health care products that are ideally suited to their particular situations (say, diabetes or heart disease), consumers might gain a great deal.[2] But other uses of algorithms should not be welcomed. If algorithms exploit a lack of information or behavioral biases on the part of identifiable people, so as to induce them to buy useless baldness cures or pointless insurance policies, those people will be harmed. We use the term "algorithmic harm" to capture this kind of injury. We catalog the different ways in which algorithms are being or may be used in consumer markets and identify the market conditions under which these uses harm consumers. We then develop legal responses that can reduce algorithmic harm.

A. Categories of Harm

The increasing use of algorithms in consumer markets gives rise to an expanding list of possible harms. We offer a taxonomy of algorithmic harms, focusing on the decision that the algorithm is asked to make. The algorithm will generally be asked to maximize profits. The question is what decisions—decisions that affect profits—are placed in the algorithm's "hands." A major set of decisions that is increasingly allocated to algorithms involves pricing. Another

[1] *See* Stephanie Assad et al., *Autonomous Algorithmic Collusion: Economic Research and Policy Implications*, 37 Oxford Rev. of Econ. Pol'y 459, 460 (2021) [hereinafter Assad (2021)] (describing how Amazon emphasizes "the possibility and the benefits of pricing automation in its marketplace with a Selling Partners API service," describing work by Chen et al. (2016) according to which "more than one third of the best-selling items on Amazon.com were priced by pricing bots in 2014/2015," quoting the European Commission's 2017 Final Report on the E-commerce Sector Inquiry that concluded: 'A majority of retailers track the online prices of competitors. Two thirds of them use software programs that autonomously adjust their own prices based on the observed prices of competitors,' and observing that "[o]ffline usage of pricing algorithms is spreading as well, for example, among gasoline retailers in northern Europe," and that "[t]here is a growing new industry of software intermediaries offering automated pricing services, from turnkey options that even small sellers can afford to fully customized pricing software for large companies. Many of these repricing companies, such as Kalibrate.com, a2i.com, and Kantify, explicitly rely on AI as a key characteristic of their algorithms."); Stephanie Assad, Robert Clark, Daniel Ershov & Lei Xu, *Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market*, CESifo Working Paper No. 8521, at *3 (2020) (describing claims on the website of "a leading algorithmic pricing software provider," according to which "their pricing software collects data and performs high frequency analysis to 'rapidly, continuously and intelligently' react to market conditions.") Popular culture offers some complicated tales of personalization and individuation, with particular reference to algorithmic harm. See, e.g., Her (2013); I'm Your Man (2021).

[2] Machine learning algorithms dubbed "recommender systems" perform tasks such as suggesting what Netflix show you may want to watch next, or what grocery item you may want to add to your Amazon Fresh cart. *See* Robin Burke, Alexander Felfernig & Mehmet H. Göker, *Recommender Systems: An Overview*, 32 AI MAGAZINE 13, 13–14 (2011); *Recommendations*, NETFLIX RESEARCH, https://research.netflix.com/research-area/recommendations; Larry Hardesty, *The history of Amazon's recommendation algorithm*, Amazon Science (Nov. 22, 2019), https://www.amazon.science/the-history-of-amazons-recommendation-algorithm. Other benefits are discussed in Assad, *supra* note 1, at 460-461 ("Algorithms guarantee faster and potentially 'better' decisions while saving costs. They are more responsive to changes in supply and demand conditions, which implies better inventory management and reduced waste, especially for perishable goods. They can also exploit consumer information, providing potentially highly personalized offers that could increase allocative efficiency. There is a general consensus that algorithmic pricing has the potential to generate significant efficiency gains and reduce transaction costs."). *See also* FEDERAL TRADE COMMISSION, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? UNDERSTANDING THE ISSUES 5–8 (2016) (listing beneficial uses of big data and algorithms) [hereinafter FTC REPORT.]

important category of decisions relates to quality, broadly understood to encompass decisions about the type of product that will be offered to a particular consumer or group of consumers. The decision can be a choice from existing items in the seller's product line or perhaps even a decision to invest in expanding the product line or shifting to a different product line.

We thus consider (1) algorithmic price discrimination, and (2) algorithmic quality discrimination (or product targeting). By discrimination we mean the setting of different prices for different consumers or the targeting of different products to different consumers. Such discrimination is fueled by individual-level data that is fed into the algorithm, e.g., the algorithm may learn that an individual consumer is a tennis fan and thus would be willing to pay a higher price for US Open tickets. We characterize the incidence of algorithmic harm for each category. To do so, we organize the analysis, for each category, into a 2x2 matrix. *See* Table 1.

	No Differentiation (Pre-Algorithmic World)	Differentiation
Perfectly Informed & Perfectly Rational Consumers	PI-PR Benchmark	PI-PR Algorithmic Harm
Imperfectly Informed or Imperfectly Rational Consumers	II-IR Benchmark	II-IR Algorithmic Harm

Table 1: General Framework for Analyzing Algorithmic Harm

The two rows distinguish between two types of consumer markets—one that is populated by perfectly informed and rational consumers (PI-PR) and another that is populated by consumers who are imperfectly informed, imperfectly rational, or both (II-IR). Of course, these are theoretical archetypes, and we are dealing with a continuum, not a dichotomy. Real-world markets are populated by a mix of more- vs. less-informed and more- vs. less rational consumers. Nevertheless, dividing the analysis in this way is useful as we explore the extent to which algorithmic harm depends on deviations from perfect information and perfect rationality. Also, as a practical matter, policymakers should be able to distinguish between markets where the majority of consumers are sophisticated (sufficiently informed and rational) and markets where the majority of consumers are unsophisticated (with significant information and rationality deficits).[3]

For each type of consumer market (i.e., for each row in Table 1), we start with the 'No Differentiation' benchmark—a pre-algorithmic world, where sellers offer the same product at the same price to everyone: medicines, clothing, laptops, food, hair loss treatments. We then compare this benchmark to a world where large data sets and sophisticated algorithms allow for at least some degree of 'Differentiation.' (In some cases, we even posit a science-fiction world of 'Full Differentiation,' where algorithms can perfectly identify each consumer's preferences and perceptions and set an individualized price or quality for every consumer. The science fiction

[3] The general problem is discussed, without reference to algorithms and algorithmic harm, in GEORGE AKERLOF & ROBERT SHILLER, PHISHING FOR PHOOLS (2016). It is worth noting that, in certain cases, the implications of imperfect information and imperfect rationality may differ. We also note that, in some parts of our analysis (e.g., Part IV), we study markets with both informed, rational consumers and imperfectly informed, and imperfectly rational consumers.

world might of course be on the way.) Our overarching conclusion will be that algorithmic differentiation is generally beneficial in PI-PR markets, but often harmful in II-IR markets.

This conclusion relates to prior work on consumer harm that predate the rise of algorithms. First, we recognize that some kinds of differentiation occurred long before machine learning algorithms were commonplace. Our claim is that the increasing use of algorithms results in higher degrees of differentiation, and we tentatively suggest that the large difference in quantity, i.e., the higher degree of differentiation, is sufficient to create a difference in quality. Our comparison between a 'No Differentiation" benchmark and a world with some (or full) differentiation helps more clearly to identify the harms caused by algorithm-enhanced differentiation. Second, the risk that uninformed, imperfectly rational consumers might be exploited by unscrupulous sellers similarly predates the rise of algorithms. Here, again, we suggest that algorithms significantly amplify the risk, e.g., by enabling the identification of specific information and rationality deficits that affect the demand of individual consumers.

B. Algorithms and Discrimination Based on Race and Sex

Our conclusion—that algorithmic harm is concentrated in II-IR markets and, more specifically, that policymakers should focus on differentiation, or discrimination, based on the consumer's information or rationality deficits—is different from that found in most prior work on algorithmic harm. This prior work has focused on the concern that algorithms will discriminate on the basis of race and sex, setting higher prices or offering inferior products to women and to members of minority groups. While acknowledging that concern, we argue that, at least in consumer markets, algorithms will often, though not always, *reduce* the risk of discrimination based on race and sex.

It follows that we should be more worried that algorithms will discriminate on the basis of information and rationality deficits, setting higher prices or offering inferior products to uninformed, biased consumers. This is not meant to suggest that algorithms will never discriminate on the basis of race and sex. Sometimes they will, and we explain when and how. The claim is only that this is not the category of harm that algorithms are most likely to exacerbate.

C. Legal Responses

The increasing use of algorithms, and the harm that such use inflicts upon imperfectly informed and imperfectly rational consumers, provide fresh support for existing efforts to reduce information and rationality deficits, especially through behaviorally-informed disclosure mandates. But our main emphasis is on two main categories of legal responses that might reduce algorithmic harm: (1) algorithmic transparency and (2) regulations policing the design and implementation of algorithms. The implementation of these regulatory responses is especially challenging, given the increasing prevalence of opaque, machine-learning algorithms. Building on recent developments in computer science and in economics, we provide suggestions for policymakers on how to open the algorithmic black-box and create meaningful transparency that can then be used to trigger market responses or regulatory scrutiny and to overcome doctrinal (*mens rea*-type) hurdles to liability for algorithmic harm. We also provide suggestions on how to police the design and implementation of these black-box algorithms, mainly through the regulatory imposition of non-discrimination constraints—including limiting any differences in outcomes experienced by imperfectly informed and imperfectly rational consumers relative to informed, rational consumers—into the algorithm's code. Our discussion of legal responses can inform policymakers in the United States and around the world who are increasingly concerned about algorithmic harm in consumer markets.[4]

***

Our focus is on algorithms deployed by sellers and service providers and the harm that they might impose on consumers. We note, however, that there are also consumer-side algorithms that can help consumers make better choices and thus mitigate the algorithmic harms that we identify. Examples include digital butlers, like Alexa, Siri and Google Assistant, that can help consumers make purchasing decisions, and more specialized apps that compare prices and help identify attractive options.[5] Without discounting the importance of consumer-side algorithms, we believe that structural asymmetries between sellers and buyers will prevent such algorithms from eliminating the harms that we identify in this Article.

The remainder of this Article is organized as follows. Parts II-IV analyze the three main categories of algorithmic harm: algorithmic price discrimination and algorithmic quality discrimination. Part V considers algorithmic discrimination based on race and sex. Part VI develops legal reform proposals to address the problem of algorithmic harm in consumer markets. Part VII concludes.

II. Algorithmic Price Discrimination: The Baseline Model

We begin with price discrimination.[6] Empirical evidence suggests that sellers are increasingly using data and algorithms to set personalized prices, i.e., to price discriminate.[7] To focus on price discrimination, we assume that the seller offers a uniform product (with uniform quality) to all consumers such that differentiation, if it occurs, is limited to the price dimension.

[4] *See*, *e.g.*, Andrew Smith, *Using Artificial Intelligence and Algorithms*, Federal Trade Commission, https://www.ftc.gov/business-guidance/blog/2020/04/using-artificial-intelligence-algorithms. *See also* Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206).

[5] *See* Michal S. Gal & Niva Elkin-Koren, *Algorithmic Consumers*, 30 HARV. J. L. & TECH. 309, 330 (2017); Marco Lippi et al., *The Force Awakens: Artificial Intelligence for Consumer Law*, 67 J. ARTIFICIAL INTELLIGENCE RES. 169 (2020). *See also* Chandra Steele, *The Best Price-Comparison Apps for Shopping*, PC Magazine (July 14, 2022) https://www.pcmag.com/picks/best-price-comparison-apps-for-shopping.

[6] The analysis in this Section is based on Oren Bar-Gill, *Algorithmic Price Discrimination When Demand Is a Function of Both Preferences and (Mis)perceptions*, 86 U. CHI. L. REV. 217 (2019).

[7] *See*, e.g., FCA, "General Insurance Pricing Practices," Financial Conduct Authority Market Study MS18/1.2 (2019); Anniko Hannak, Gary Soeller, David Lazer, Alan Mislove & Christo Wilson, *Measuring Price Discrimination and Steering on E-Commerce Web Sites*, Proceedings of the 2014 Conference on Internet Measurement, at 305-318 (2014); Ipsos, London Economics and Deloitte Consortium, "Consumer Market Study on Online Market Segmentation Through Personalised Pricing/Offers in the European Union," European Commission Report (2018); ARIEL EZRACHI & MAURICE STUCKE, VIRTUAL COMPETITION 89–96 (2016); Jakub Mikians et al., *Detecting Price and Search Discrimination on the Internet*, Proceedings of the 11th Association for Computing Machinery Workshop on Hot Topics in Networks, Oct 2012, at *4-5; Sal Thomas, *Does Dynamic Pricing Risk Turning Personalisation into Discrimination?*, CAMPAIGN (Oct 22, 2014), https://www.campaignlive.co.uk/article/does-dynamic-pricing-risk-turning-personalisation-discrimination/1317995; *In-Store Tracking Tech Gets Personalized*, PYMNTS.COM (Feb 9, 2018), https://www.pymnts.com/news/retail/2018/adobe-retail-apocalypse-personalization-brick-and-mortar/.

Price discrimination requires some degree of market power.[8] For expositional simplicity, we focus on the extreme, monopoly case.[9]

A. PI-PR Markets

We first consider the effect of algorithmic pricing in markets with informed, rational consumers. To do so, we first derive the no differentiation, PI-PR Benchmark, and then compare this benchmark to the outcome with full differentiation, thus identifying the PI-PR Algorithmic Harm. The PI-PR Benchmark is presented in Figure 1, using the most basic market setup with a linear, downward sloping demand curve and a linear, horizontal supply curve (reflecting a fixed-per-unit-cost assumption; let k denote the per-unit cost).[10] The intersection of the demand curve with the supply curve, at (Q_C, P_C), represents the perfect-competition equilibrium, where Q_C represents the equilibrium quantity and P_C represents the equilibrium price (which is equal to the per-unit cost, k).[11] But, as explained above, we focus on the monopoly case. Compared to the perfect-competition case, a monopolist will set a higher price, $P_M > P_C$, and sell fewer units of the product, $Q_M < Q_C$.

Consumer surplus is represented by the red triangle; it is equal to the difference between the consumer's willingness to pay (WTP) and the price, P_M, aggregated across all consumers. Some consumers have a high WTP. They are represented by the high points on the left side of the demand curve, and they enjoy more surplus. Other consumers have a lower WTP. They are represented by the lower points of the demand curve, close to Q_M, and they enjoy less surplus. The seller's surplus is represented by the blue rectangle and is equal to the number of units sold multiplied by the difference between the monopoly price and the per-unit cost: $Q_M \cdot (P_M - k)$. Social welfare is, by definition, equal to the sum of the consumer surplus and the producer's (monopolist's) surplus. The black triangle represents the monopoly deadweight loss: Because of the higher price that the monopolist charges, consumers who should buy the product refrain from purchasing it (specifically, the lost quantity is given by $Q_C - Q_M$); and the welfare that these lost purchases would have produced constitutes the monopoly deadweight loss.

[8] *See, e.g.*, Lars A. Stole, *Price Discrimination and Competition*, *in* 3 HANDBOOK OF INDUSTRIAL ORGANIZATION 2221, 2226 (Mark Armstrong & Robert H. Porter eds., 2007) ("It is well known that price discrimination is only feasible under certain conditions: (i) firms have short-run market power, (ii) consumers can be segmented either directly or indirectly, and (iii) arbitrage across differently priced goods is infeasible.")

[9] For an analysis of an oligopoly case, see Oren Bar-Gill, *Consumer Misperceptions in a Hotelling Model: With and Without Price Discrimination*, 176 J. INST. & THEORETICAL ECON. 180 (2020).

[10] *See, e.g.*, Andrew Mas-Colell, Michael D. Whinston & Jerry R. Green, MICROECONOMIC THEORY 321 (4th ed. 2012); HAL R. VARIAN, INTERMEDIATE MICROECONOMICS: A MODERN APPROACH 292–94 (8th ed. 2010).

[11] *See, e.g.*, Mas-Colell, *supra* note 9, at 316–322; Varian, *supra* note 9.

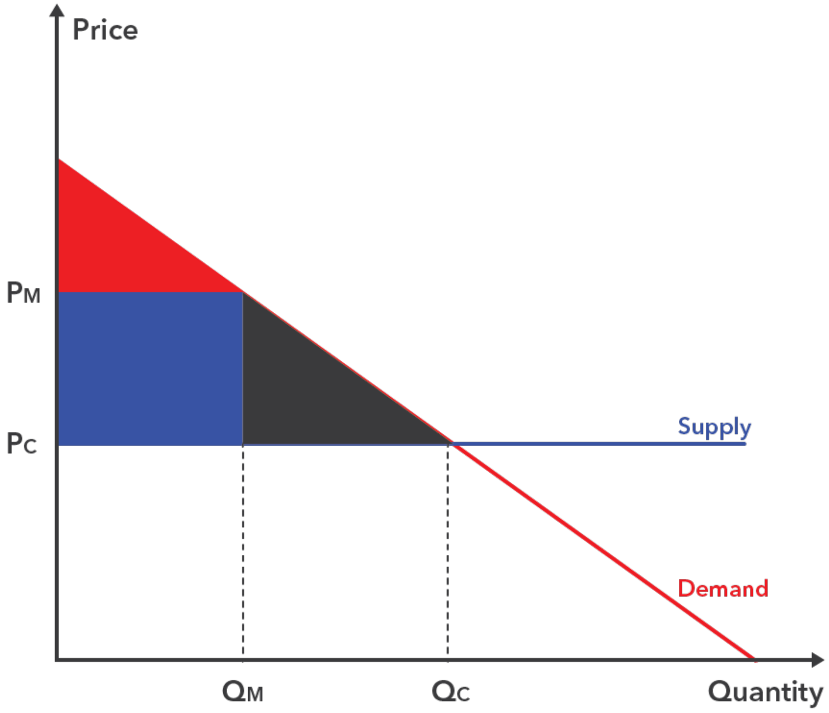


Figure 1: The 'No Differentiation,' PI-PR Benchmark

Next, we consider the 'full differentiation' outcome, where the monopolist charges each consumer a different, personalized price.[12] *See* Figure 2. Using big data and sophisticated algorithms, the monopolist will identify each consumer's WTP and set a personalized price just below this WTP. Thus, a consumer with a high WTP on the left side of the demand curve will pay a high price; and a consumer with a lower WTP towards the middle or right side of the demand curve will pay a lower price. The seller's surplus is represented by the blue rectangle and is equal to the difference between the consumer's WTP and the per-unit cost, k, aggregated across all consumers. A price discriminating monopolist keeps the entire surplus to itself; there is no consumer surplus. Observe that the quantity sold is Q_C, as in the competition case. Price discrimination allows the monopolist to increase the quantity sold – from Q_M to Q_C – thus eliminating the deadweight loss and increasing overall social welfare. However, this efficiency gain comes at a steep distributional price; the entire surplus goes to the monopolist and consumers

[12] Partial differentiation, or imperfect price discrimination, is considered in an extension. *See* Sec. III.D. below. Welfare effects can be non-monotonic in the degree of differentiation, such that consumers (especially poor consumers) benefit from a move from no differentiation to partial differentiation but are them harmed by a move from partial differentiation to full differentiation. Still, the comparison between no differentiation and full differentiation—here and in Sec. II.B. (on II-IR markets)—is useful in demonstrating our main argument: a higher degree of differentiation is more harmful to consumers in II-IR markets, as compared to PI-PR markets; and this higher degree of differentiation may or may not increase efficiency in II-IR markets, whereas it generally increases efficiency in PI-PR markets. This comparison between the effect of differentiation in PI-PR markets and in II-IR markets extends to the partial differentiation case, at least with linear demand curves. *See* Oren Bar-Gill, *Price discrimination with consumer misperception*, 28 Applied Economics Letters 829 (2021).

are left with nothing.[13] Still, the efficiency gain is worth emphasizing. It is a powerful argument in favor of price discrimination in markets with informed, rational consumers.

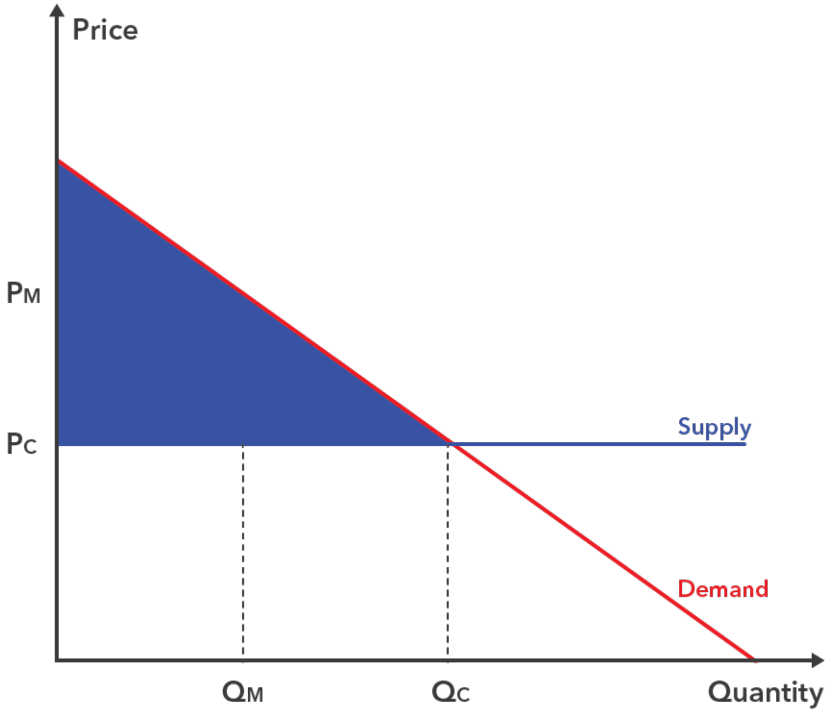


Figure 2: 'Full Differentiation' in PI-PR Markets

B. II-IR Markets

We now consider the effect of algorithmic pricing in markets where consumers are imperfectly informed or imperfectly rational (or both). To do so, we first derive the no differentiation, II-IR Benchmark, and then compare this benchmark to the outcome with full differentiation, thus identifying the II-IR Algorithmic Harm.

Before proceeding, we must consider how imperfect information and imperfect rationality manifest in our analytical-graphical framework. These imperfections affect consumers' WTP. A consumer who overestimates the benefit from the product will have a higher WTP, and a consumer who underestimates the benefit from the product will have a lower WTP. We begin with overestimation, which is probably the more prevalent problem (as sellers have an incentive to

[13] There are additional distributional implications for the consumer side of the market. Consumers with a higher WTP, who would have purchased the product at the (no discrimination) monopoly price, suffer an affirmative loss, as they pay more for the same product. Consumers with a lower WTP are not affected – without price discrimination they would have been priced out of the market, and with price discrimination they still get a zero (net) surplus. Budget constraints and wealth effects (*see infra* note 14) add nuance to this analysis: If WTP is smaller than the preference-based benefit, then perfect price discrimination affirmatively helps consumers with a low WTP while hurting consumers with a high WTP. If WTP is positively correlated with wealth, then perfect price discrimination results in progressive redistribution.

promote overestimation and fight underestimation); the underestimation case is discussed in an extension. We initially assume that the degree of overestimation is not correlated with consumers' preference-based WTP, namely, that the average bias level is the same for consumers with a higher preference-based WTP at the left-hand side of the demand curve and for consumers with a lower preference-based WTP towards the middle and right-hand side of the demand curve. (This assumption is relaxed in Section C below.) Now, in addition to the actual demand curve, we have a perceived demand curve. In Figures 3 and 4, the actual demand curve is represented by the solid red line, and the perceived demand curve is represented by the dashed red line.

The II-IR Benchmark is presented in Figure 3. In the PI-PR Benchmark, the monopoly price was determined by the demand curve.[14] In the II-IR Benchmark, the price is determined by the perceived demand curve. Therefore, the monopoly price with misperception, P_M', is higher than the monopoly price without misperception, P_M. The quantity sold with misperception, Q_M', is also higher than the quantity sold without misperception, Q_M.[15] Turning to welfare: The higher price reduces the actual consumer surplus, which is represented by the red triangle. (More precisely, the red triangle represents transactions that create positive consumer surplus; to see the full consumer surplus, we need to subtract transactions that create negative consumer surplus, as described below.) The *perceived* surplus is larger—the perceived extra surplus is represented by the light-red trapezoid. Overestimation causes some consumers to purchase the product even though its actual value to them is lower than the price, P_M'.

The loss incurred by these consumers is represented by the purple triangle. This loss reduces the (actual) consumer surplus. Indeed, the consumer surplus might be negative—the purple triangle might be larger than the red triangle. But whatever consumers lose, the monopolist gains. The purple triangle is part of the blue rectangle, which represents the monopolist's surplus. Therefore, we have a distributional effect, but no reduction in efficiency. Indeed, misperception increases efficiency. By inflating demand, the overestimation bias increases the quantity sold—from Q_M to Q_M'—and thus reduces the monopoly deadweight loss, which is represented by the black triangle. Notice that the black triangle in Figure 3 is smaller than the black triangle in Figure 1.[16]

[14] *See, e.g.*, Mas-Colell, *supra* note 10 at 384–86; Varian, *supra* note 10, at 441–43.

[15] The overestimation inflates demand and thus increases the quantity sold. The higher price somewhat tempers this quantity-increasing effect, but cannot reverse it.

[16] When the misperception is even stronger and the perceived demand curve shifts even higher above the actual demand curve, the quantity, Q_M', can be larger than Q_C. In this case, the black triangle disappears entirely, and the problem of insufficient purchases is replaced with a problem of excessive purchases. Specifically, consumers in the $[Q_C, Q_M']$ range inefficiently purchase the product. Misperception can either increase or decrease overall efficiency in this market, depending on the relative magnitudes of the insufficient purchases problem (without misperception) and the excessive purchases problem (with misperception).

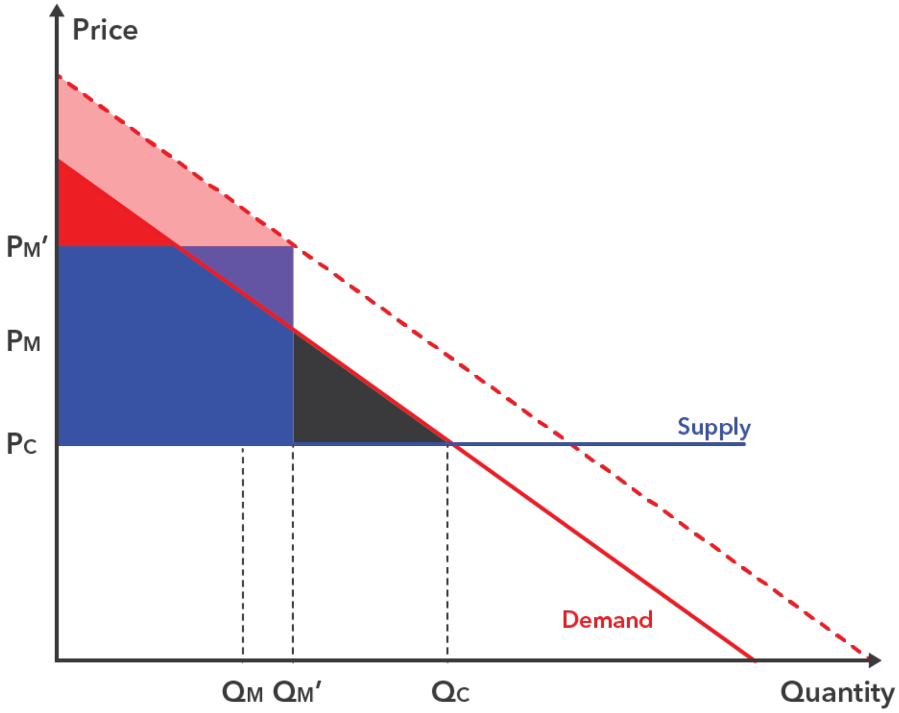


Figure 3: The ‘No Differentiation,’ II-IR Benchmark

Next, we consider the ‘full differentiation’ outcome, where the monopolist charges each consumer a different, personalized price, equal to the consumer’s WTP. See Figure 4. Whereas WTP derived only from preferences in the PI-PR case, now WTP is a product of both preferences and misperceptions. Price discrimination allows the monopolist to “march down” the demand curve, setting different prices for different consumers. In the PI-PR case, the monopolist marched down the *actual* demand curve. In the II-IR case, the monopolist is marching down the *perceived* demand curve. Turning to welfare, in the PI-PR case the monopolist extracted the entire surplus. Consumers gained nothing, but also lost nothing. In the II-IR case, the monopolist is also extracting perceived surplus, which is represented by the purple trapezoid. This extra gain to the monopolist is a loss to consumers; the purple trapezoid represents a transfer from consumers to the monopolist—a distributional effect with no efficiency implications.[17] But there are also efficiency implications. Consumers in the $[Q_C, Q_C{}']$ range should not purchase the product. They buy only

[17] There are additional distributional implications for the consumer side of the market, especially if we add budget constraints and wealth effects (see *supra* note 13): Consumers with a high WTP, who would have purchased the product and gained a positive surplus in the absence of price discrimination, lose that positive surplus and more. Consumers with a low WTP, who would have been priced out of the market in the absence of price discrimination, now purchase the product and pay a price equal to their full WTP, including both the preference-based and misperception-based components. But, as noted above, without misperceptions budget constraints may keep the WTP below the preference-based level; and with misperceptions, if the bias level is sufficiently low, the full WTP may still be below the preference-based level. Therefore, consumers with a low WTP may actually benefit from price discrimination. And if WTP is positively correlated with wealth, then perfect price discrimination may result in progressive redistribution. Still, as compared to the parallel effect in the standard model (see supra note ???), misperception reduces the magnitude of any progressive redistribution.

because of the misperception, because they overestimate the product's value. These purchases create an efficiency loss, which is borne entirely by consumers. This loss is represented by the light-red triangle below the supply curve.

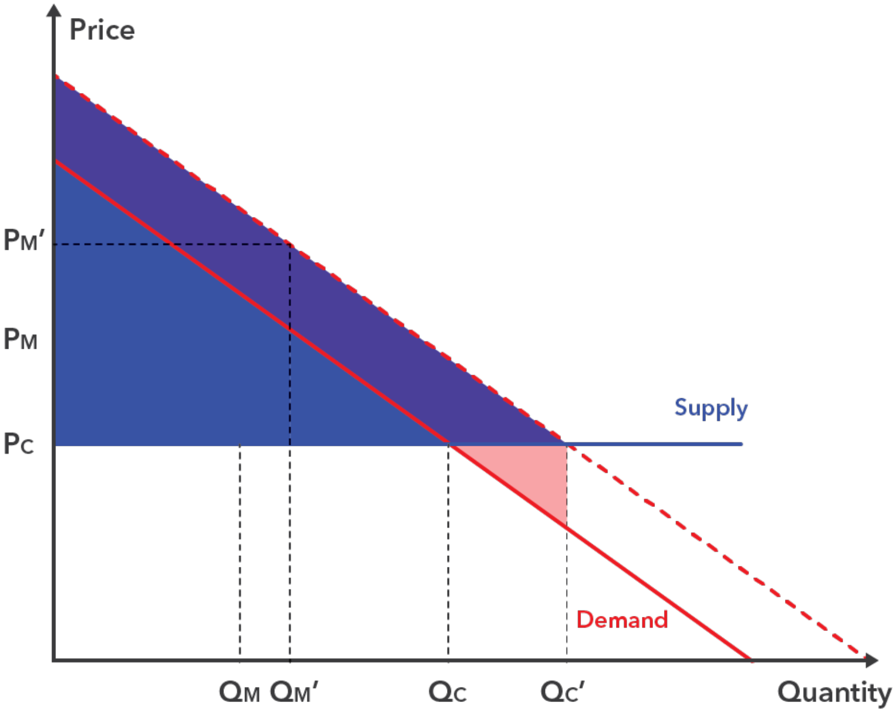


Figure 4: 'Full Differentiation' in II-IR Markets

In the PI-PR case, where WTP is derived from preferences alone (*see supra* Section III.A), price discrimination hurts consumers but increases efficiency. Specifically, consumers enjoy no surplus at all, but deadweight loss is eliminated. In the II-IR case, price discrimination hurts consumers even more and may either increase or decrease efficiency. Consumers are hurt more because now they give up surplus that they do not have—perceived surplus—and thus end up with a loss. In terms of efficiency, the insufficient quantity problem is avoided, but an excessive quantity problem is created. Whether price discrimination increases or decreases efficiency depends on the relative magnitudes of the black triangle in Figure 3 and the light-red triangle in Figure 4.[18]

C. Summary

In the PI-PR case, algorithms increase efficiency by eliminating the monopoly deadweight loss. At the same time, they harm consumers by erasing the consumer surplus. In the II-IR case, algorithms harm consumers even more—not only do they erase the consumer surplus, but they also create a negative consumer surplus by setting prices above the consumer's actual benefit. In

[18] When the misperception is stronger such that $Q_M{}'$ is larger than Q_C, price discrimination definitely decreases efficiency. In this case, there is an excessive quantity problem even in the absence of price discrimination, and price discrimination only exacerbates this problem.

addition, the algorithm-enabled price discrimination might reduce rather than increase efficiency in the II-IR case.

It is important to note that the algorithm does not set out to harm consumers; it is programmed to maximize profit. To maximize profit, the algorithm seeks out consumers' WTP for different products and services. The extent and nature of the resulting algorithmic harm depend on different factors that determine the WTP. In particular, a consumer's WTP depends on (1) preferences—a consumer will pay more for a product that generates a greater benefit in terms of preference satisfaction, broadly understood, (2) wealth (or budget constraints)—a rich consumer will be able (and willing) to pay more than a poor consumer, and (3) misperceptions—a consumer who overestimates the benefit from a product, because of some information or rationality deficit, will pay more for that product. An algorithm, designed to maximize profit, cares only about the bottom-line WTP, not about the factors that influence the WTP. But the harm that this algorithm causes very much depends on these underlying factors. As we have seen, when WTP is largely determined by preferences and wealth (the PI-PR case), the algorithm causes limited harm and may even generate socially desirable outcomes. It is when WTP is significantly influenced by misperceptions (the II-IR case) that algorithms raise particular concern.

III. Algorithmic Price Discrimination: Extensions

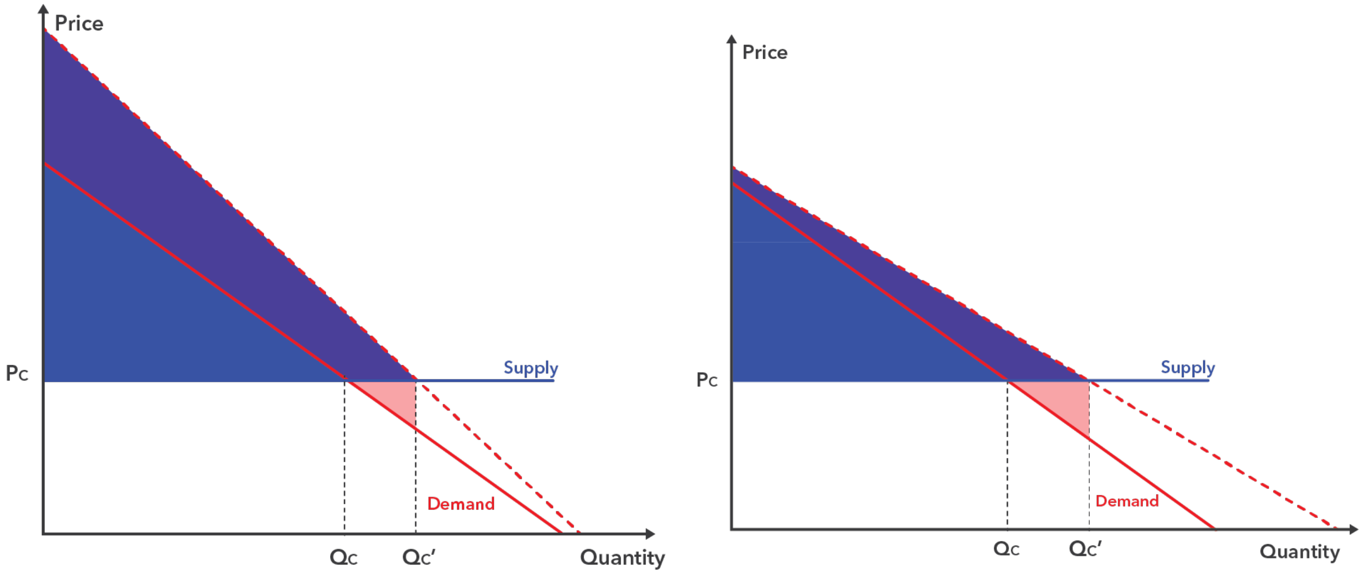
A. Misperceptions that are Correlated with the Preference-based WTP

Our baseline analysis above assumed that the degree of misperception is not correlated with the consumer's preference-based WTP. Graphically, this assumption was represented by a perceived demand curve that was parallel to the actual demand curve. Put differently, the perceived demand curve was represented by an upward shift from the actual demand curve. If there is a positive correlation between the degree of misperception and preference-based WTP, then the distance between the perceived and actual demand curves is larger at the left-hand side of the graph and smaller at the right-hand side of the graph. *See* Figure 5(a). Conversely, if there is a negative correlation between the degree of misperception and preference-based WTP, then the distance between the perceived and actual demand curves is smaller at the left-hand side of the graph and larger at the right-hand side of the graph. *See* Figure 5(b).[19]

Correlation between the degree of misperception and the preference-based WTP adds nuance to the normative assessment of algorithmic price discrimination. The extra harm that consumers incur is larger in the positive correlation case and smaller in the negative correlation case. The positive per-unit production cost, i.e., the Supply Curve, truncates the perceived demand curve and the overestimation bias, and thus the consumer harm from overpayment. This truncation effect is smaller in the positive correlation case and larger in the negative correlation case. Moreover, if higher preference-based WTP (at the left-hand side of the demand curve) represent

[19] The correlation between consumers' bias levels and their preference-based WTP will be positive when bias is proportional to (actual) value. The correlation between consumers' bias levels and their preference-based WTP will be negative when bias is negatively correlated with wealth. It is not that poor people are more prone to bias; rather, rich people can afford to hire expert advisers—human or virtual—that mitigate bias and misperception. And so, if preference-based WTP is positively correlated with wealth, and wealth is negatively correlated with bias levels, then the preference-based WTP will be negatively correlated with bias levels. *See* Bar-Gill, *supra* note 6, at 246.

richer consumers and lower preference-based WTP (at the right-hand side of the demand curve) represent poorer consumers, then richer consumers incur relatively larger harm in the positive correlation case and poorer consumers incur relatively larger harm in the negative correlation case. In terms of efficiency, the cost of price discrimination is measured by the welfare-reducing transactions that are entered into by overestimating consumers (represented by the red triangles in Figure 5)—a cost that needs to be compared to the monopoly deadweight loss in the absence of price discrimination. Price discrimination is more likely to reduce efficiency, when the cost from the welfare-reducing transactions is higher (i.e., when the red triangle is larger). In the positive correlation case, the welfare loss from inefficient transaction is higher when the per-unit production cost is high; in the negative correlation case, the loss is higher when the per-unit product cost is low.



(a) Positive Correlation

(b) Negative Correlation

Figure 5: Correlated Misperceptions

B. Underestimation

Our baseline analysis assumed that consumers overestimate the benefit from a product or service. But in some markets, we can expect underestimation. For example, consumers likely underestimate the benefit from health insurance (e.g., because they underestimate future healthcare costs). And present-biased consumers likely underestimate the benefit from a more fuel-efficient car. What are the welfare implications of algorithmic price discrimination when consumers underestimate the benefit?

Starting with the no-discrimination benchmark: Underestimation reduces the price that the monopolist sets (since the monopoly price is determined by the demand curve, which is pushed down by the misperception). Underestimation also reduces the quantity sold.[20] Turning to welfare: In the PI-PR case, without misperception, monopoly pricing prevents some efficient purchases, thus creating the infamous monopoly deadweight loss. The underestimation bias prevents additional, efficient purchases from taking place, thus increasing the deadweight loss.[21]

Now add (perfect) price discrimination: The monopolist charges an individual price for each consumer, based on each consumer's WTP. Turning to welfare, price discrimination clearly increases efficiency – it reduces the deadweight loss, i.e., more consumers purchase the product. The effect on the consumer surplus, however, is ambiguous. In the PI-PR case the monopolist extracted the entire surplus. Consumers gained nothing. Here the monopolist can extract only the underestimated perceived surplus. The consumers are left with the difference between the actual surplus and the perceived surplus. So consumers enjoy a positive surplus, but it is not clear whether this surplus is larger or smaller than the surplus that they enjoy in the absence of price discrimination. On the one hand, more consumers buy the product and enjoy this difference between the actual and perceived surplus. On the other hand, the consumers who would have purchased the product also in the absence of price discrimination enjoy a smaller surplus (because they are charged a higher, personalized price).

To conclude: In the PI-PR case, price discrimination hurts consumers but increases efficiency. Specifically, consumers enjoy no surplus at all, but there is no deadweight loss. With overestimation, price discrimination hurts consumers even more and may either increase or decrease efficiency. Here, with underestimation, price discrimination clearly increases efficiency and may or may not hurt consumers. Therefore, algorithmic price discrimination is less worrisome, and thus legal intervention is less needed in markets with underestimation.

C. Behavior-Based Pricing

We now consider behavior-based pricing (BBP), where the algorithm discriminates based on the consumer's past behavior.[22] To clarify, our baseline analysis of algorithmic price discrimination did not specify the source of the WTP information that the algorithm used to price discriminate; and the baseline analysis applies to situations where the WTP information is based on the consumer's past behavior. But when sellers' information about consumers' WTP is based on the consumers past purchasing decisions, there are additional welfare effects to consider. First, the welfare analysis now includes a dynamic component: over time, as sellers and their algorithms accumulate more information about consumers' past behavior, the degree of price discrimination increases. Second, in PI-PR markets consumers will strategically adjust their purchasing behavior in earlier periods in order to obtain lower prices in later periods. Such strategic response mitigates, and may even eliminate, algorithmic harm from BBP in PI-PR markets.[23] As before, the harm to

[20] The underestimation deflates demand and thus decreases the quantity sold. The lower price somewhat tempers this quantity-decreasing effect, but cannot reverse it.

[21] Consumers who purchase the product despite the misperception enjoy a larger surplus, thanks to the lower price.

[22] The analysis of this extension draws on the excellent discussion in Haggai Porat, *Consumer Protection and Disclosure Rules in the Age of Algorithmic Behavior-Based Pricing* (Apr. 7, 2022) (unpublished manuscript) (on file with author).

[23] Recent work has begun to develop algorithms that anticipate strategic responses and are robust to such responses. *See* Daniel Björkegren, Joshua E. Blumenstock & Samsun Knight, *Manipulation-proof Machine Learning*, arXiv preprint arXiv:2004.03865 (2020). These algorithms would be expected to increase sellers profits and reduce the consumer surplus in PI-PR markets (i.e., where sophisticated consumers are likely to respond strategically to BBP).

consumers is concentrated in II-IR markets, where many consumers are unaware of the algorithm's BBP and do not respond strategically. (Indeed, in the BBP extension, we define PI-PR markets as those where most consumers are aware of the seller's BBP and respond strategically, and we define II-IR markets as those where most consumers are unaware of the seller's BBP and thus do not respond strategically.)

BBP pricing is already practiced in certain consumer markets and its prevalence is likely to increase. Amazon experimented with BBP in 2000, setting higher prices for consumers who purchased certain DVDs.[24] More recently, Uber has been accused of engaging in BBP,[25] but there is no clear proof. And the Airline Tariff Publishing Company (ATPCO), which is co-owned by several large airlines, announced in October 2019 that it is developing a dynamic pricing tool, which can adjust pricing based on consumers' prior transactions.[26] Finally, it is quite clear that large tech companies, like Amazon and Apple collect data on consumers' purchasing behavior; and that data aggregators collect and sell similar data to smaller companies.[27] It would be surprising if these data are not fed into sellers' pricing algorithms.

To illustrate the equilibrium outcomes and welfare implications of algorithmic BBP, we consider a simple two-period model. In the earlier period, the (monopolist) seller has limited information and thus sets a single price for all potential customers. In the later period, the seller sets two prices—a higher price for consumers who purchased in the earlier period and a lower price for those who did not. Suppose, for example, that in the earlier period Uber sets a single price for all potential riders. Uber then observes that a certain consumer declined a ride at this price. The Uber algorithm will identify this consumer as a low-WTP consumer and offer her a lower price in the later period. In contrast, another consumer who took the ride in the earlier period will be identified as a high-WTP consumer and offered a higher price in the later period.

To ascertain the effect of algorithmic BBP, we begin with the pre-algorithmic benchmark. In this pre-algorithmic world, a monopolistic seller will set the same single (monopoly) price in both the early and late periods. This means that the same, higher-WTP consumers purchase the product in both periods; and the same, lower-WTP consumers are excluded from the market in both periods. With algorithmic BBP, in the earlier period fewer consumers will purchase the product. In II-IR markets this is because the seller will increase the early-period price (relative to the pre-algorithmic benchmark), in order to more effectively discriminate between low- and high-

[24] Amazon stopped these experiments when consumers found out about them and expressed their unhappiness. *See Amazon pricing flap*, CNNMONEY (Sept. 28, 2000), https://money.cnn.com/2000/09/28/technology/amazon/.

[25] *See* Arwa Mahdawi, *Is your friend getting a cheaper Uber fare than you are*, THE GUARDIAN (Apr. 13, 2018), https://www.theguardian.com/commentisfree/2018/apr/13/uber-lyft-prices-personalized-data.

[26] *See* Barbara Peterson, *Airline Dynamic Pricing Getting Closer to Reality, Says ATPCO*, TRAVEL MARKET REPORT, Oct. 1, 2019.

[27] *See* Kai Hao Yang, *Selling Consumer Data for Profit: Optimal Market-Segmentation Design and Its Consequences*, 112 Am. Econ. Rev. 1364, 1365 (2022) ("[D]ata companies such as Acxiom and Datalogix gather and sell personal information including government records, financial activities, online activities, and medical records to retailers."); Kirsten Martin, *Data Aggregators, Consumer Data, and Responsibility Online: Who Is Tracking Consumers Online and Should They Stop?*, 32 The Info. Soc'y 51, 57 (2016) ("Broad data aggregators summarize information across diverse contexts into profiles and sell aggregated information to companies looking for a specific, target market."); Federal Trade Commission, Data Brokers: A Call for Transparency and Accountability 13, 23 (2014) ("[D]ata brokers obtain detailed, transaction-specific data about purchases from retailers and catalog companies" and turn them into marketing products that "enable the data brokers' clients to create tailored marketing messages to consumers.")

WTP consumers in the later period. In PI-PR markets this is because a group of strategic consumers will not purchase the product even though they value it more than the charged price. For example, a consumer can strategically decline the Uber ride offer, even if the benefit from the ride exceeds the offered price, in order to secure lower price offers in the future. In the later period more consumers purchase the product under BBP. Specifically, low-WTP consumers who were excluded from the market in the earlier period will face a lower price in the later period and thus enter the market.

In terms of consumer surplus, in both PI-PR markets and II-IR markets consumers with lower WTP, who are likely poorer, benefit from BBP, because they face a lower price and thus can enter the market even if only in the later period, whereas they are excluded from the market in both periods without BBP. The main difference between PI-PR markets and II-IR markets is with respect to consumers with higher WTP who are likely richer. In II-IR markets, these consumers are harmed by BBP, because they now face a higher price in the later period. In PI-PR markets, these consumers also pay a higher price in the later period. But they pay a lower price in the earlier period, because sellers reduce the earlier-period price to limit the number of consumers who strategically refrain from purchasing. Across both periods, consumers with higher WTP benefit from BBP in PI-PR markets. Therefore, algorithmic BBP increases the overall consumer surplus in PI-PR markets.[28] In contrast, algorithmic BBP reduces the overall consumer surplus in II-IR markets, as the harm to the higher-WTP consumers exceeds the benefit to the lower-WTP consumers. Still, if our social welfare function places greater weight on lower-WTP consumers who are likely poorer, then BBP can be desirable, or at least less undesirable, even in II-IR markets. In any event, we see, once again, that concern about algorithmic harm should be smaller in PI-PR markets and greater in II-IR markets.[29]

D. Additional Extensions

Misperception about the price. In important consumer markets—think mortgages, credit cards, cellular services, broadband, insurance—pricing is complex and multidimensional. In these markets, the main concern is about price misperception, namely, that consumers might not fully understand the pricing structure and thus underestimate the overall price that they will end up paying for the product or service. Consumers might not pay attention to certain components of the pricing structure; some of those components might be in some sense shrouded or not highly visible. Or consumers might underestimate the probability of triggering a certain price dimension, such as a late fee on a credit card or mortgage. When algorithms can be used to identify and exploit such price misperceptions, consumers will incur harm that is similar to the harm analyzed above.

[28] In PI-PR markets, BBP helps consumers and harms sellers. Therefore, in the early period, sellers would prefer to commit to refrain from using BBP, if they could. But such a commitment may well prove impossible: in the later period, armed with reams of data and the algorithms to analyze it, sellers will have a strong incentive to engage in BBP; and sophisticated consumers will anticipate this in the early period and respond accordingly.

[29] In terms of total surplus, in both PI-PR markets and II-IR markets BBP increases the total number of transactions across the two periods, i.e., the increase in the number of later-period transactions outweighs the decrease in earlier-period transactions, and thus the total surplus increases. In the Appendix, we provide a formal analysis of algorithmic BBP and derive its implications for both total surplus and consumer surplus (including implications for different subgroups of consumers), in PI-PR markets and in II-IR markets.

Indeed, the effects of price underestimation are analytically identical to the effects of value overestimation that we analyzed above.

Discriminating between more- and less-sophisticated consumers. For analytical purposes, we distinguished between PI-PR markets on the hand and II-IR markets on the other hand. But we have also noted that in practice, most markets include both more- and less-sophisticated consumers. In these markets, sellers will employ algorithms to try to discriminate between these groups of consumers—offering better deals to the more sophisticated consumers and bad deals to the less sophisticated consumers. For example, sellers can offer generally high-priced products with a few good deals hidden among their offerings. More sophisticated consumers will find those good deals,[30] whereas less-sophisticated consumers will not.[31]

Imperfect price discrimination. For expositional purposes, we compare no price discrimination to perfect price discrimination. In reality, the degree of price discrimination is continuous, and algorithms are likely to affect a shift towards higher degrees of price discrimination, but still falling short of perfect price discrimination. In PI-PR markets, increased, yet imperfect price discrimination can be attractive: the rich pay more (because they have a higher WTP) and the poor pay less (because they have a lower WTP). If the rich pay more than the poor for (say) electricity, food, and automobiles, there are gains in terms of both efficiency and fair distribution.[32] In II-IR markets, even with imperfect price discrimination, there is a risk that the poor, and the rich, will end up paying more than their preference-based WTP. Our basic result—that a higher degree of price discrimination is more harmful to consumers in II-IR markets, and may or may not increase efficiency in such markets (as compared to PI-PR markets where it generally increases efficiency)—extends to the imperfect price discrimination case, with linear demand curves.[33]

Competition. As explained above, some degree of market power is a precondition for price discrimination and, for simplicity, we analyzed a monopolistic market. How would the analysis change if sellers, while enjoying some market power, are still subject to the forces of competition? On the one hand, competition might force algorithmic harm, as sellers who fail to utilize algorithms will lose out to competitors who do.[34] On the other hand, competition can reduce algorithmic harm

[30] For example, by using the RECAP tools described in CASS R. SUNSTEIN & RICHARD H. THALER, NUDGE (2008). *See also* Emir Kamenica, Sendhil Mullainathan & Richard Thaler, *Helping Consumers Know Themselves*, 101 AM. ECON. REV. PAPERS & PROCS. 417, 417–18 (2011).

[31] These ideas were expressed in David Laibson's comments on Kamenica, Mullainathan & Thaler, *id.*, at the American Economics Association Annual Meeting in 2011. The welfare implications of such discrimination depends on what sellers will do if they cannot discriminate—will they offer better deal to everyone or the worse deal to everyone?

[32] *See*, *e.g.*, Jean-Pierre Dube & Sanjog Misra, *Personalized Pricing and Consumer Welfare*, J. Pol. Econ. (forthcoming) (finding, in a field experiment, that while personalized pricing reduces the overall consumer surplus, many consumers, with lower WTP, benefit from lower prices). More generally, the economics literature, which has been focused on PI-PR markets, shows that the effect of price discrimination on consumer welfare is ambiguous. *See*, *e.g.*, Eeva Mauring, *Search and Price Discrimination Online*, CEPR Discussion Paper 15729 (2022) (finding that, with rational consumers and WTP based on preferences and budget constraint (but not misperceptions), regulation that limits price discrimination can help consumers, but only if it is strict enough.)

[33] *See* Oren Bar-Gill, *Price discrimination with consumer misperception*, 28 Applied Economics Letters 829 (2021). Relaxing the linear-demand assumption can lead to more nuanced results.

[34] *See* OREN BAR-GILL, SEDUCTION BY CONTRACT 16 (2012) (arguing that sellers who fail to exploit consumer biases will lose out to competitors who do).

by constraining sellers' ability to engage in price discrimination. In addition, one seller might reveal algorithmic abuses by her competitor in attempt to win over consumers.

Outside options. The preceding analysis assumed a monopoly seller, such that the only outside option was 'no purchase.' If we relax the monopoly assumption, WTP can be influenced by the consumer's actual and perceived outside options. For example, if a consumer can purchase the product from Seller #1 at a price of $100, i.e., if the consumer has an "outside option" of getting the product for $100, then her WTP, when facing Seller #2, will never exceed $100. Some consumers have access to multiple, competing sellers. These consumers will have a lower WTP. Other consumers do not have access to competing sellers (e.g., because they don't have a car, or don't have internet access, or don't have a bank account). These consumers will have a higher WTP.

Algorithms will be able to identify consumers with fewer, or less attractive, outside options and offer them higher prices. And, like other factors that influence the WTP, the existence and features of outside options might be subject to misperception. Specifically, an imperfectly informed and imperfectly rational consumer might underestimate her outside options (e.g., she might underestimate her ability to get a lower price from a competing seller). As a result, the consumer will have a higher WTP. An algorithm trained to track WTP would set a higher price for this consumer, even if the consumer could in fact get a lower price from a competing seller.[35]

Consumers will have a higher WTP if they do not have attractive outside options. Consumers will also have a higher WTP if they are less likely to shop around and explore their outside options, perhaps because they are very busy or less savvy. Once again, algorithms will identify such consumers and charge them a higher price.

Cost-based price discrimination. We have thus far focused on situations where the cost to the seller of providing the good or the service does not depend on the consumer's characteristics and where the algorithmic pricing tracks the consumer's WTP. But there are also important situations where the seller's cost depends on the consumer's characteristics and the algorithm tracks these cost-affecting characteristics, setting higher prices for higher-cost consumers and lower prices for lower-cost consumers. Consumer credit markets are perhaps the leading example. When a lender offers a loan to a borrower, the cost to the lender of this loan depends on the likelihood that the borrower will repay the loan. When the probability of repayment is higher, the cost to the lender is lower and thus the lender can offer a lower price, i.e., a lower interest rate. And when the probability of repayment is lower, the cost to the lender is higher and thus the lender will set a higher interest rate. The pricing algorithms thus tracks borrower characteristics that predict the probability of repayment, such as income, debt overhang and the consumer's history of debt repayment across multiple lenders. This is what credit scoring models do, and these models are increasingly algorithm-based.[36]

[35] *See* Simon Jäger, Christopher Roth, Nina Roussille & Benjamin Schoefer, *Worker Beliefs About Outside Options* 3 (Nat'l Bureau of Econ. Rsch., Working Paper No. 29623, 2021).

[36] *See* Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow & Lyn C. Thomas, *Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research*, 247 Eur. J. of Operational Rsch. 124, 124 (2015) ("[I]n application scoring, . . . lenders employ predictive models, called scorecards, to estimate how likely an applicant is to default," which are "routinely developed using classification algorithms."); Lyn C. Thomas, Consumer

When algorithmic pricing tracks cost, rather than WTP, the concern about algorithmic harms is diminished. It is less objectionable for sellers or lenders to charge higher prices when they face higher costs. As before, in assessing the welfare implications of algorithmic pricing a comparison to the pre-algorithmic world is instructive. If lenders cannot distinguish between low-risk and high-risk borrowers, then they would set a single interest rate that reflects average risk. Low-risk borrowers would then cross-subsidize high-risk borrowers, creating both winners (high-risk borrowers) and losers (low-risk borrowers). If high-risk borrowers are generally poorer, then this pre-algorithmic outcome can be distributionally attractive, and pricing algorithms that eliminate the cross-subsidization would then be socially harmful. But it is also possible that, in the pre-algorithmic world, low-risk borrowers would reject the single interest rate and exit the market. Realizing that only high-risk borrowers remain, lenders would then increase the interest rate. There would be no cross-subsidization, only a smaller market serving only high-risk borrowers. If this is the pre-algorithmic benchmark, then algorithmic pricing would increase welfare by expanding the market to low-risk borrowers.

It is important to emphasize that our leading distinction between PI-PR markets and II-IR markets is less important when price discrimination tracks cost, or risk, rather than WTP. WTP is a consumer-side feature that is commonly influenced by the consumer's imperfect information and imperfect rationality. Consumers often overestimate the benefits of a product, resulting in an inflated WTP. In contrast, cost is a seller-side feature, even though it is influenced by certain consumer characteristics. When algorithms allow sellers to adjust the price so that it matches the cost of serving the particular consumer, the consumer's imperfect information and imperfect rationality do not enter the equation (at least not directly). Therefore, the welfare analysis of algorithmic cost-based pricing is similar in both PI-PR markets and in II-IR markets.

IV. Algorithmic Quality Discrimination

In this Part, we shift our focus from price discrimination to quality discrimination. To focus on quality discrimination, we assume that the seller offers a uniform price to all consumers, such that differentiation, if it occurs, is limited to the quality dimension. The algorithm matches different consumers with different product designs.

Credit Models: Pricing, Profit and Portfolios 25–26 (2009) (characteristics that affect credit score "can include socio-economic characteristics, like the age, residential status, and employment of an individual, their past credit performance including late or missed payments, and their existing debt obligations"); Bee Wah Yap, Seng Huat Ong & Nor Huselina Mohamed Husain, *Using Data Mining to Improve Assessment of Credit Worthiness via Credit Scoring Models*, 38 Expert Sys. with Applications 13274, 13274 (2011) (while "Linear Discriminant Analysis and logistic regression are two popular statistical tools to construct credit scoring models," the advance of data mining provide new "predictive modeling and classification techniques such as decision tree, neural networks, support vector machine (SVM), and k-nearest neighbors" to improve credit scoring models); Yilun Jin *et al.*, *A Novel Multi-Stage Ensemble Model With a Hybrid Genetic Algorithm for Credit Scoring on Imbalanced Data*, 9 IEEE Access 143593, 143594 (2021) ("Recently, the advancements in artificial intelligence, such as ensemble learning-based methods, evolution algorithm-based methods, and clustering technique-based methods have been used in credit scoring fields.")

A. PI-PR Markets

In the PI-PR case, this type of algorithmic discrimination can be welfare enhancing. Consider the market for laptops and assume, for simplicity, that there are two types of laptops. The first has a large, super-high-definition screen and a powerful graphics card. The second has a lower-end screen and graphics card, but a super-fast Central Processing Unit (CPU) and extra Random Access Memory (RAM). It would be welfare enhancing if the algorithm were to offer the first laptop to a graphic designer and the second to a computer scientist who needs to analyze large datasets. In more extreme cases, every consumer could be offered the specific laptop that is most likely to fit her particular needs.[37]

The examples could easily be proliferated. The central point is that in light of the immense diversity of both preferences and products, a great deal might be gained in terms of welfare if an algorithm could help to "match" particular desires and needs with particular offerings. So long as we are dealing with PI-PR cases, there are welfare gains if the matches are accurate. To be sure, we would have little need for the assistance of an algorithm if search costs were zero; in that case, people could find the right product. A key advantage of the algorithm, under the circumstances we are assuming, is that it eliminates search costs.

B. II-IR Markets

When studying algorithmic quality discrimination, the interesting II-IR case is one where some consumers are informed and rational, but others are imperfectly informed or imperfectly rational. In this case, algorithmic quality discrimination might be welfare reducing. Specifically, if imperfectly informed or imperfectly rational consumers overestimate the benefit from a lower-quality product, mistakenly preferring this product over an objectively superior product, the algorithm would offer the superior product to the informed, rational consumers while offering the lower-quality product to the imperfectly informed or imperfectly rational consumers. This algorithmic outcome is harmful if, in a pre-algorithmic world with no quality discrimination, sellers would offer the superior product to *all* consumers.

In these scenarios, one group of consumers is offered lower-quality products, rather than just different-quality products (as in the laptop example from Section A above). But algorithmic quality discrimination can also help consumers in II-IR markets. For instance, if a sufficiently large number of imperfectly informed or imperfectly rational consumers *under*estimate the benefit from a higher-quality product, mistakenly preferring a lower-quality product, then in a pre-algorithmic world all consumers would be offered the lower-quality product; whereas algorithmic quality discrimination allows the seller to offer the higher-quality product at least to the informed, rational consumers.[38]

[37] The idea is that, with positive search costs, the product that is offered or prioritized by the seller is more likely to be purchased by the consumer, even if the consumer could potentially find an alternative—not offered or not prioritized—product. This is a case where the label "PI" (= Perfect Information) is imprecise, since it can be taken to imply costless search.

[38] *Cf.* Jeannie Marie Paterson, Shanton Chang, Marc Cheong, Chris Culnane & Suelette Dreyfus, *The Hidden Harms of Targeted Advertising*, 9 INT'L J. CONSUMER L. & PRAC. 1, 8 (2021).

Consider a market with two products, P1 and P2. To focus on the effect of benefit and perceived benefit, we assume that the cost, to Seller, of manufacturing the two products is identical, and for expositional purposes we let this cost be zero. Consumers enjoy a benefit of 200 from P1 and a smaller benefit, 100, from P2. We assume that some of the consumers are informed and rational, and thus accurately identify the benefits that they would derive from each product: 200 from P1 and 100 from P2, while others are imperfectly informed or imperfectly rational, and thus misperceives the benefit from one of the products. We distinguish between the case where the lower, P2 benefit is overestimated and the case where the higher, P1 benefit is underestimated. Market power is such that Seller gets half of the perceived surplus and the consumer gets half of the perceived surplus. (Note that, since the cost is zero, half of the perceived surplus is equal to half of the perceived benefit.) This equal division of the perceived surplus is achieved by setting the price equal to half of the perceived benefit.

1. Overestimation

Consider the following examples:

Example 1a: There are two types of cars: (i) a larger car with more leg-room and a bigger trunk (P1), which provides a benefit of 200; and (ii) a smaller that comes with a higher-end entertainment system (P2), which provides a benefit of 100. One-half (1/2) of consumers are informed and rational, and thus accurately identify the benefits that they would derive from each car. The other half (1/2) of consumers overestimate the number of hours that they will spend listening to opera in the car and thus overestimate the benefit from P2, mistakenly thinking that it is 300 (rather than 100).[39]

Example 1b: Same as Example 1a, except that one-quarter (1/4) of consumers are informed and rational, and the other three-quarters (3/4) overestimate the benefit from P2, mistakenly thinking that it is 300 (rather than 100).

In a world with big data and sophisticated algorithms, Seller can distinguish between the biased and unbiased consumers. Therefore, Seller will offer the larger vehicle to the unbiased consumers, at a price of 100 (which is half of the actual benefit, 200). And Seller will offer the smaller car with the high-end entertainment system to the biased consumers who overestimate the benefit from the entertainment system, at a price of 150 (which is half of the perceived benefit, 300). In an algorithmic world, in Example 1a, Seller's overall profit is: $\frac{1}{2} \times 100 + \frac{1}{2} \times 150 = 125$; and the overall consumer surplus is: $\frac{1}{2} \times (200 - 100) + \frac{1}{2} \times (100 - 150) = 25$. And, in Example 1b, Seller's overall profit is: $\frac{1}{4} \times 100 + \frac{3}{4} \times 150 = 137.5$; and the overall consumer surplus is: $\frac{1}{4} \times (200 - 100) + \frac{3}{4} \times (100 - 150) = -12.5$.

To appreciate the potential algorithmic harm in such cases, we must compare the quality-discrimination outcome to the no-differentiation benchmark. What would car sellers do in a pre-algorithmic world, where they cannot distinguish the biased consumers from the unbiased

[39] To focus on situations where the overestimation bias is potentially most troubling, we assume that the overestimated benefit from P2 exceeds the accurately perceived benefit from P1, i.e., that the bias flips the relative desirability of the two products.

consumers? Unable to discriminate, the sellers would offer the same car to all consumers.[40] But which car will they offer? Would they offer the larger car or the smaller? And what price will they set? If Seller offers P1, then misperception doesn't play a role (since only the benefit from P2 is misperceived). Seller sets a price 100 and earns a profit of 100. Note that all consumers buy P1. If Seller offers P2, then Seller would forgo the business generated by the unbiased consumers and set a price of 150, at which only overestimators would make the purchase. Seller's profit would then be $\frac{1}{2} \times 150 = 75$ in Example 1a and $\frac{3}{4} \times 150 = 112.5$ in Example 1b, reflecting a higher per-unit profit but a smaller number of units sold.[41] Since 75 < 100, in Example 1a Seller will offer the larger car to all consumers, and the consumer surplus will be 200 – 100 = 100. And, since 112.5 > 100, in Example 1b Seller will offer the smaller car at a price that will attract only the biased consumers, and the consumer surplus will be $\frac{3}{4} \times (100 - 150) = -37.5$.

To assess the welfare effects of algorithmic quality discrimination, we compare the pre- and post-algorithmic worlds. In Example 1a, quality discrimination harms consumers who enjoy a surplus of 100 in the pre-algorithmic world and only 25 in the post-algorithmic world. In a pre-algorithmic world, all consumers get the superior product (the larger car), P1, whereas in the post-algorithmic world, the biased consumers get the inferior product (the smaller car), P2, and overpay for it. In contrast, in Example 1b, quality discrimination helps consumers who lose 37.5 in the pre-algorithmic world and lose only 12.5 in the post-algorithmic world. In a pre-algorithmic world, unbiased consumers are left out of the market, whereas in the post-algorithmic world they get the larger care, P1, and their purchases increase the overall consumer surplus. (In both worlds, biased consumers get the smaller care, P2, and overpay for it.)

2. Underestimation

Consider the following examples:

Example 2a: There are two types of cars: (i) a highly fuel-efficient hybrid vehicle (P1), which provides a benefit of 300; and (ii) a gas guzzler but one that comes with fancier seats and a higher-end entertainment system (P2) and provides a benefit of 200. One-half (1/2) of consumers are informed and rational, and thus accurately identify the benefits that they would derive from each car. The other half (1/2) of consumers are present biased and thus underestimate the fuel-efficiency advantage of P1; these consumers mistakenly thinking that the benefit from P1 is 100 (rather than 300).[42]

[40] We compare the option of offering only the larger vehicle or offering only the smaller vehicle. But there is another possibility: If sellers cannot discriminate, they might offer a third product design (i.e., not one of the two product designs described in the text). In this case, algorithmic discrimination might help some consumers while harming others.

[41] Seller will never offer P2 at a price that will attract all consumers. Intuitively, in order to sell the smaller car to all consumers, Seller would have to reduce the price to a level that even unbiased consumers would be willing to pay. But if such a low price is needed to capture the entire market with the smaller car, it is more profitable for Seller to capture the entire market with the larger car that can fetch a higher price.

[42] To focus on situations where the underestimation bias is potentially most troubling, we assume that the underestimated benefit from P1 is lower than the accurately perceived benefit from P2, i.e., that the bias flips the relative desirability of the two products.

Example 2b: Same as Example 2a, except that three-quarters (3/4) of consumers are informed and rational, and the other one-quarter (1/4) underestimate the benefit from P2, mistakenly thinking that it is 100 (rather than 300).

In a world with big data and sophisticated algorithms, Seller can distinguish between the biased and unbiased consumers. Therefore, Seller will offer the hybrid vehicle to the unbiased consumers, at a price of 150 (which is half of the actual benefit, 300). And Seller will offer the gas guzzler with the fancy seats and the high-end entertainment system to the biased consumers who underestimate the fuel-efficiency advantage of the hybrid car, at a price of 100 (which is half of the actual benefit, 200). In an algorithmic world, in Example 2a, Seller's overall profit is: $\frac{1}{2} \times 150 + \frac{1}{2} \times 100 = 125$; and the overall consumer surplus is: $\frac{1}{2} \times (300 - 150) + \frac{1}{2} \times (200 - 100) = 125$. And, in Example 2b, Seller's overall profit is: $\frac{3}{4} \times 150 + \frac{1}{4} \times 100 = 137.5$; and the overall consumer surplus is: $\frac{3}{4} \times (300 - 150) + \frac{1}{4} \times (200 - 100) = 137.5$.

To appreciate the potential algorithmic harm in such cases, we must compare the quality-discrimination outcome to the no-differentiation benchmark—to a pre-algorithmic world where, unable to distinguish between the biased and unbiased consumers, sellers must offer the same car to all consumers. If Seller offers the gas guzzler (P2), then misperception doesn't play a role (since only the benefit from P1 is misperceived). Seller sets a price 100 and earns a profit of 100. Note that all consumers buy P2. If Seller offers the hybrid vehicle (P1), then Seller would forgo the business generated by the biased consumers and set a price of 150, at which only unbiased consumers would make the purchase. Seller's profit would then be $\frac{1}{2} \times 150 = 75$ in Example 2a and $\frac{3}{4} \times 150 = 112.5$ in Example 2b, reflecting a higher per-unit profit but a smaller number of units sold.[43] Since 75 < 100, in Example 2a Seller will offer the gas guzzler to all consumers, and the consumer surplus will be 200 – 100 = 100. And, since 112.5 > 100, in Example 2b Seller will offer the hybrid car at a price that will attract only the unbiased consumers, and the consumer surplus will be $\frac{3}{4} \times (300 - 150) = 112.5$.

To assess the welfare effects of algorithmic quality discrimination, we compare the pre- and post-algorithmic worlds. In Example 2a, quality discrimination helps consumers who enjoy a surplus of 112.5 in the pre-algorithmic world and a higher surplus of 125 in the post-algorithmic world. In a pre-algorithmic world, all consumers get the inferior product (the gas guzzler), P2, whereas in the post-algorithmic world, the unbiased consumers get the better product (the hybrid), P1. Also, in Example 2b, quality discrimination helps consumers who enjoy a surplus of 112.5 in the pre-algorithmic world and a higher surplus of 137.5 in the post-algorithmic world. In a pre-algorithmic world, biased consumers are left out of the market, whereas in the post-algorithmic world they at least get the gas guzzler, P2 (which still provides a positive benefit).

[43] Seller will never offer P1 at a price that will attract all consumers. Intuitively, in order to sell the hybrid car to all consumers, Seller would have to reduce the price to a level that even biased consumers would be willing to pay. But if such a low price is needed to capture the entire market with the hybrid car, it is more profitable for Seller to capture the entire market with the gas guzzler that can fetch a higher price.

C. Summary

In the PI-PR case, algorithmic quality discrimination is welfare enhancing, as it allows for a better matching between products and consumers. In the II-IR case, the picture is more complicated. When some consumers overestimate the benefit from an inferior product, algorithmic quality discrimination harms consumers if the superior product would have been offered to all consumers in a pre-algorithmic, no-differentiation world. If the inferior product would have been offered at an inflated price only to the biased consumers, then algorithmic quality discrimination helps the unbiased consumers (and does not harm the biased consumers). When some consumers underestimate the benefit from a superior product, algorithmic quality discrimination helps consumers because, in a pre-algorithmic, no-differentiation world either (i) the inferior product would have been offered to all consumers; or (ii) the superior product would have been offered at a price that completely excludes biased consumers from the market (whereas algorithmic discrimination allows biased consumers to at least get the inferior product).

V. Algorithmic Discrimination Based on Race and Sex

We now turn to an issue that is receiving a great deal of attention: algorithmic discrimination on the basis of race and sex. We argue that the increasing use of algorithms need not exacerbate that problem, and may even help to reduce it. Algorithms programmed to maximize profits are less likely to engage in statistical discrimination or taste-based discrimination, and they are unlikely to suffer from the unconscious bias that afflicts many human decisionmakers. To the extent that algorithms still discriminate on the basis of race and sex, it would often (not always) be easier to police algorithmic discrimination, as compared to discrimination by human decisionmakers. For these reasons, we argue for a broadening of focus—supplementing attention to algorithmic discrimination based on race and sex with algorithmic discrimination based on information and rationality deficits, as manifested in the algorithmic harms that we analyzed in the preceding Parts of this Article.

After providing some background on antidiscrimination law in Section A, we elaborate in Section B on the benefits that algorithms present in the context of race-based and sex-based discrimination. While we argue that algorithms may reduce the incidence of discrimination, we also emphasize that algorithms may sometimes discriminate on the basis of race and sex. In Section C, we discuss precisely when such discrimination is most likely to occur.

A. Background: Antidiscrimination Law

To understand the problems introduced by algorithms, it is important to lay out the fundamentals of antidiscrimination law, which has long been focused on two different problems. The first is disparate treatment; the second is disparate impact.[44] If we are concerned about the possibility that algorithms might promote discrimination, or on the contrary reduce it, we need to distinguish sharply between the two. The Constitution, and all civil rights laws, forbid disparate

[44] For an overview, see Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 672, 694 (2016).

treatment.[45] The Constitution does not concern itself with disparate impact,[46] but some civil rights statutes do.[47]

The prohibition on disparate treatment reflects a commitment to a kind of neutrality. When the prohibition is in place, favoring men over women, or whites over Blacks, is essentially forbidden. When it occurs, discrimination might be a product of "taste" or "animus." A seller might prefer, personally, not to sell to Blacks. Or the seller might have no particular racial preference, but might believe, or know, that her employees prefer not to sell to Blacks, or to work with Hispanics. Alternatively, disparate treatment might be a product of statistical discrimination. For example, a seller might believe that women generally have a higher WTP than men, or a lender might believe that Blacks are more likely to default on their loans as compared to whites.

The prohibition on disparate impact means, in brief, that if some requirement or practice has a disproportionate adverse effect on members of specified groups (Blacks, women), the requirement or practice must be shown to be adequately justified.[48] Suppose, for example, that a police department establishes a height requirement for its employees. If this practice has a disproportionate adverse effect on women, the practice will be invalidated unless the department can show that the practice is justified by "business necessity,"[49] e.g., that the height requirement is an essential filter for police department employees, given the nature of the job.

B. Algorithmic Benefits

Algorithms are less likely to engage in both taste-based and statistical discrimination, and they are unlikely to suffer from the unconscious bias that afflicts many human decisionmakers. In a world without algorithms, we might well observe a significant amount of racism and sexism, producing taste-based discrimination. Algorithms do not have tastes, and unless they are designed to discriminate on the basis of race and sex, they will not do so. In a world without algorithms, we might also observe a significant amount of statistical discrimination, in which race and sex are used as proxies for relevant characteristics, such as willingness to pay, ability to repay, and so forth.

One goal of civil rights laws is to forbid that form of discrimination (as an instance of disparate treatment), but let us stipulate that those laws are imperfectly enforced, which means that statistical discrimination will occur. Now let us suppose that algorithms are able to make fine-grained judgments, based on rich data, about who is willing or able to pay more for a product or service and who is more or less likely to repay a loan. If so, algorithms that are programmed to maximize profits should not be expected to engage in race- or sex-based statistical discrimination.

[45] *See, e.g.*, *Washington v. Davis*, 426 U.S. 229, 238 (1976); *Personnel Adm'r of Mass. v. Feeney*, 442 U.S. 256, 281 (1979). The Constitution is understood to forbid disparate treatment along a variety of specified grounds, including race and sex. In extreme cases, the existence of disparate treatment is obvious because a facially discriminatory practice or rule can be shown to be in place). In other cases, no such practice or rule can be identified, and the question is whether a facially neutral practice or rule was motivated by a discriminatory purpose.
[46] *See Washington*, 426 U.S. at 238.
[47] *See, e.g.*, *Griggs v. Duke Power Co.*, 401 U.S. 424, 434–35 (1971) (interpreting Title VII of the 1964 Civil Rights Act).
[48] *See Griggs*, 401 U.S. at 436.
[49] 42 U.S.C. §§ 2000e-2(k)(1)(A)–(B).

The reason is that if they can make fine-grained judgments, they would not need to rely on proxies, which are likely to be unnecessarily coarse.

Suppose, for example, that women are less likely to repay loans than men are, and that human decisionmakers take that point into account in deciding on interest rates for loans. Algorithms ought to be able to use far less crude approaches; they should not use sex as a proxy.[50] Crude proxies of that kind are unlikely to be excellent predictors, and algorithms should be expected to use excellent predictors. For example, an algorithm tasked with predicting the likelihood of loan repayment would use data on the borrowers past loans, rent payments, utility payments, and a host of other factors that are statistically correlated with repayment patterns. Similarly, the use of algorithms will reduce the effects of unconscious bias. The preceding example assumed that the borrower's sex is in fact correlated with repayment probability. But it may well be that there is no such correlation; only the lender who suffers from unconscious bias mistakenly believes that a correlation exists. A shift to algorithmic loan pricing would avoid the adverse implications of the unconscious bias.

The case of taste-based discrimination can be analyzed similarly. Algorithms will focus on the relevant characteristics of consumers. If John has a credit record identical to Joan's, John and Joan will be treated similarly, and if existing evidence suggests that John is willing to pay more than Joan, it will not matter that John is male and that Joan is female. Again, algorithms do not have tastes, and they will not show taste-based discrimination unless they have been programmed to do so or they learn that accommodating the discriminatory tastes of some group helps to maximize their assigned objective (e.g., profit maximization). For this reason, we could easily imagine situations in which the use of algorithms is likely to have particular benefits for (say) women and people of color, as compared to a situation in which decisions are made by human beings.[51]

Since algorithms are less likely to engage in taste-based or statistical discrimination and are likely to avoid the adverse implications of unconscious bias, the rise of algorithms in consumer markets may be beneficial from the perspective of race- and sex-based discrimination. This does not mean that algorithms will not discriminate on the basis of race and sex. Indeed, as explained below, there are circumstances where algorithms might exacerbate discrimination on the basis of race and sex. But even when algorithms discriminate, there is a potential benefit: it will often be easier to detect discrimination by algorithms, as compared to discrimination by human decisionmakers, if the law appropriately adjusts to the rise of algorithms. We discuss such adjustment in Part VI below.

[50] More precisely, algorithms should not use sex as a single or dominant proxy. Kleinberg et al. show that algorithms should use both the neutral data and the data on sex, as this would achieve superior accuracy *and* less sex-based discrimination. The reason is that if sex is excluded as an input, the algorithm will mis-rank women among themselves (formally, because various features, such as age, when interacted with sex, have different effects on outcome prediction, such that excluding sex forces the algorithm to use the same measure of the effect of age for both sexes, mis-ranking within each of the groups). *See* Kleinberg et al., Algorithmic Fairness (2018).

[51] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, J. L. ANALYSIS 113, 114 (2018).

C. Algorithmic Harms

Thus far, then, the problem of race- and sex-based discrimination seems more serious for human beings than for algorithms. But that conclusion is far too simple and in important contexts, it might be wrong. Suppose, for example, that the data on which algorithms are trained reflects human bias. If loan performance records reflect human judgments that are themselves discriminatory, and if algorithms take account of such records, then they will discriminate.

There is also the question of disparate impact. Even if the algorithm is programmed to exclude race and sex data, the algorithm will pick up other variables (or combinations of other variables) that are closely correlated with race or sex. Suppose, for example, that people of color are less likely to have graduated from college than are white people, or that people of color are less likely to have good credit ratings than white people. If an algorithm that is programmed to maximize profits identifies a correlation between these variables and profits, and treats consumers in accordance with them, it will produce a disparate impact on people of color. It might be challenging, of course, to know whether there is a disparate impact, and to test the question whether it might be justified under prevailing standards. Moreover, if there is disparate impact, it is not always clear that the disparate impact is harmful, e.g., if race is correlated with income, people of color may be offered lower prices. We will return to these issues.[52]

In some circumstances, algorithms are likely to produce such disparate impacts even though they would not occur in a pre-algorithmic world. The reason is that algorithms might well have access to information that human beings lack. For example, in a pre-algorithmic world, when a consumer makes a purchase by phone or online, the seller would not know whether the consumer is white or African American. But an algorithm, armed with endless data linking the consumer's phone number or IP address to a host of traits and past behaviors, might pick up variables (or combinations of variables) that are correlated with race. Also, even if a human decisionmaker observes the consumer's race, she might not know that race is correlated with a higher WTP, or with a lower ability to repay, and thus might offer the same price, or interest rate, to both Black and white consumers. An algorithmic decisionmaker, on the other hand, will know that race-based pricing maximizes profits and thus discriminate between the Black and white consumers. (As noted above, even if the algorithm is programmed not to consider race directly, it will learn to discriminate based on other variables that are correlated with race.)

With respect to our concerns here, algorithms might discriminate even in the PI-PR case, where WTP or ability to repay is correlated with race or sex. But some of the most serious problems will arise if members of traditionally disadvantaged groups are especially vulnerable to imperfect information and imperfect rationality, or in other words, if the distinction between the PI-PR case and the II-IR case is itself correlated with race or sex. One possibility is that past discrimination might have resulted in limited access to information and to mechanisms, such as expert advice, that can mitigate bias. If that is so, the particular harms identified in Parts II-IV above would disproportionally fall on traditionally disadvantaged groups. We might end up with cases of disparate impact.[53]

[52] *See* discussion *infra* in Part VI.C.

[53] *Cf.* Paterson, *supra* note 38, at 8–9.

A final note: In Part III, we considered the possibility that consumers would respond strategically to algorithmic behavior-based pricing (BBP), e.g., by declining a value-increasing period 1 purchase in order to elicit a lower period 2 price. Since such strategic responses reduce sellers' profits, an algorithm might learn to rely on less accurate but more immutable characteristics, like race and sex, which are largely immune to strategic behavior on the part of consumers. While this theoretical possibility should be acknowledged in PI-PR markets, we believe that most consumers lack the level of sophistication needed to respond strategically to BBP in a way that would push algorithms to rely on rough proxies like race and sex. Put differently, for present purposes, most markets are likely II-IR markets.

VI. Legal Reforms

To reduce algorithmic harm in consumer markets, we propose three sets of legal reforms. The first would use, and expand on, current initiatives designed to increase information and to reduce the impact of behavioral biases, with an understanding that the rise of algorithms imposes fresh threats to consumer welfare. Though these initiatives are not new, the argument on their behalf is significantly strengthened by an appreciation of the dangers that we have explored.

The second set of reforms would involve a right to algorithmic transparency, designed to ensure that consumers (and others) can know about the nature, uses, and consequences of algorithms. The central idea here is that sunlight might serve as a disinfectant, reducing the incidence and magnitude of algorithmic harm. The argument for algorithmic transparency will be divided into an easy case and a hard case. The easy case applies to white-box algorithms, where the programmer pre-defines how inputs are combined to generate outputs. Here, the transparency reforms require that firms share what they know about the algorithms that they use. The hard case applies to black-box algorithms—machine-learning algorithms, where the process of manipulating inputs to generate outputs is opaque, even to the programmer.

With machine-learning algorithms, the challenge is in opening the black-box, i.e., in creating previously unavailable knowledge about how algorithmic decisions are made. Only then can we talk about the transparent sharing of this knowledge. Building on recent developments in computer science and in economics, we will provide suggestions for policymakers on how to "open" the black-box and "interpret" the algorithm's decision-making process. The process of knowledge creation can be performed by the firms themselves and reported to regulators, or by regulators based on data and code supplied by the firms. Transparency about how algorithmic decisions are made may trigger a public or market reaction. It may also trigger regulatory scrutiny. Specifically, by forcing firms to learn how their algorithms actually work, this reform would open the door to liability under legal doctrines that require knowledge or intent.

The third reform would involve more direct regulation of the design and implementation of algorithms that are used in consumer markets, mainly through the regulatory imposition of non-discrimination constraints—including limiting disparate impacts on imperfectly informed and imperfectly rational consumers—into the algorithm's code. Also, in appropriate cases, regulators should consider prohibitions and bans on the use of algorithms to harm consumers in the ways that we have discussed.

A. Less I, More P

Because algorithmic harm is a product of II-IR situations, the most obvious remedies involve consumer protection in the form of information disclosure and debiasing, designed to move II-IR situations in the direction of PI-PR situations. In federal law, disclosure policies are of course pervasive,[54] and the need for those policies increases to the extent that algorithms can be used to exploit ignorance and behavioral biases. Many existing disclosure policies are explicitly meant to overcome such biases,[55] and some of them are behaviorally informed, in the sense that they are based on an understanding of specific biases, and attempt to design a remedy that will reduce the risk that some seller (human or algorithmic) will exploit them.

For example, some such policies are directed against hidden fees and hence to counteract limited attention.[56] Other disclosure policies, such as graphic health warnings, can be seen as an effort to counteract unrealistic optimism.[57] Present bias can be a special problem in the context of both health and savings,[58] and creative efforts have been made to overcome that bias on the part of savers.[59] While behaviorally-informed disclosure policies show promise in some contexts, their efficacy in other contexts is quite limited.[60] The central point is that behaviorally informed disclosure policies, seeking to counteract biases, will have increasing importance to the extent that algorithms, employed in consumer markets, can exploit these biases.[61]

B. A Right to Algorithmic Transparency? Easy Case: White-Box Algorithms

It would be unrealistic to think that efforts to provide information and to counteract behavioral biases can entirely eliminate algorithmic harm. A more targeted disclosure policy would require transparency *with respect to the nature, the uses, and the consequences of algorithms in the relevant markets*.[62] In various areas of regulatory law, transparency of certain kinds is mandatory,[63] largely on the theory that sunlight can be a kind of disinfectant, helping consumers to make better choices and potentially deterring certain practices.[64] As explained above,

[54] For discussion, *see generally* OREN BAR-GILL, SEDUCTION BY CONTRACT (2012).

[55] *See* PRESS RELEASE, U.S. DEP'T OF TRANSP., EPA, DOT UNVEIL THE NEXT GENERATION OF FUEL ECONOMY LABELS (Aug. 1, 2019)., https://www.transportation.gov/briefing-room/epa-dot-unveil-next-generation-fuel-economy-labels#:~:text=Fuel%20Economy%3A%20The%20label%20shows,in%20a%20gallon%20of%20gasoline.

[56] *See* Sumit Agrawal, Souphala Chomsisengphet, Neale Mahoney & Johannes Stroebel, *A Simple Framework for Estimating Consumer Benefits from Regulating Hidden Fees*, 43 J. L. STUD. S239, S240 (2014).

[57] *See* 21 C.F.R. 1141 (2021) (imposing labeling requirements for cigarette packages and advertisements).

[58] *See* Yang Wang & Frank A. Sloan, *Present bias and health*, 57 J. RISK & UNCERTAINTY 177, 178 (2018).

[59] *See* Hal E. Hershfield, *Future self-continuity: how conceptions of the future self transform intertemporal choice*, 2011 ANN. N.Y. ACAD. SCI. 30, 31 (2013).

[60] *See*, e.g., Sumit Agarwal et al., *Regulating Consumer Financial Products: Evidence from Credit Cards*, 130 Q. J. Econ. 111 (2015) (finding that a CARD Act disclosure failed to reduce overall interest payments). For a general critique of disclosure regulation, *see* OMRI BEN-SHAHAR & CARL SCHNEIDER, MORE THAN YOU WANTED TO KNOW: THE FAILURE OF MANDATED DISCLOSURE (2014).

[61] Efforts to counteract "dark patterns" have particular importance, because algorithms might promote actions that fall squarely in that category (such as default terms and hidden fees). *See* Jamie Luguri & Lior Jacob Strahilevitz, *Shining a Light on Dark Patterns*, 13 J. L. ANALYSIS 43, 44, 47, 61 (2021).

[62] *Cf.* Paterson et al., *supra* note 38, at 13–14 (2021).

[63] 14 C.F.R. Part 399. *See also* EU General Data Protection Regulation (GDPR), articles 13-15 (mandating that "meaningful information about the logic" of automated systems be made available to data subjects).

[64] For example, the Centers for Medicare & Medicaid Services has attempted to increase hospital price transparency, with the goal of enabling consumers to shop and compare prices across hospitals and estimate the cost of care before

we focus here on a requirement that firms share information about their algorithms that they already have. This type of disclosure requirements is generally appropriate when firms use white-box algorithms, namely, algorithms that implement a set of instructions that is specified by the firm—by the seller or by the firm that wrote the algorithm. A "right to algorithmic transparency" can be designed to uncover and mitigate the kinds of practices that concern us here.

Start with algorithmic price discrimination. Suppose, for example, that a seller's algorithm divides consumers into four categories corresponding to their income and wealth; suppose too that wealthier consumers are charged higher prices. Companies might have to disclose that (not particularly alarming) fact. Or suppose that an algorithm is told to use data on a consumer's borrowing and saving behavior to identify myopic consumers (who tend to borrow more and save less) and then offer such consumers low introductory prices and high long-term prices.[65] Or suppose that the algorithm is told to identify consumers who would likely overestimate the benefit from the firm's product, and then set higher prices for these consumers. A transparency requirement would force firms to disclose that their algorithms are searching for myopic consumers, or for consumers who suffer from an overestimation bias, and offering different prices to these consumers. It is easily imaginable that transparency could deter some of the practices on which we have focused here.[66]

Next, consider algorithmic quality discrimination. Suppose that companies use algorithms to identify less sophisticated consumers and offer them inferior products. Or to elaborate on the prior example, suppose that an algorithm is told to identify myopic consumers and then offer these consumers products and prices with immediate benefits and deferred costs, such as a gas-guzzling car or a cheap printer with expensive ink. Transparency could deter such harmful targeting.

Finally, consider the case of discrimination on the basis of race or sex. Antidiscrimination law clearly prohibits algorithms that are designed to identify women or racial minorities and single them out for disparate treatment. A transparency requirement could help enforce this prohibition. Such a requirement could also deter attempts to skirt the antidiscrimination law. Suppose that a seller, in attempt to avoid liability, designs the algorithm to ignore direct data on a consumer's sex, and to use the consumer's height instead (knowing that height is correlated with sex). If the seller is forced to disclose the role that height plays in its algorithmic decisionmaking, then the seller may be deterred from using height where this physical characteristic is clearly used as a proxy for

going to the hospital. CENTERS FOR MEDICARE & MEDICAID SERVICES, *Hospital Price Transparency* (Dec. 1, 2021). And the Department of Transportation has issued a number of rules designed to increase price transparency, so as to enable consumers to have more clarity about what they are buying or not buying, and to discourage certain kinds of fees. Supp. Notice of Proposed Rulemaking, *Transparency of Airline Ancillary Service Fees*, 14 C.F.R. Part 399 (Jan. 17, 2017).

[65] It is not clear that a high level of borrowing and a low level of savings necessarily implies myopia or present bias; it could also imply exponential discounting with a high discount rate, i.e., it could imply a preference rather than a bias. In that case, low introductory prices and high long-term prices can be welfare increasing.

[66] Under the Fair Credit Reporting Act (FCRA), 15 U.S.C. §§ 1681–1681x, when a company denies a customer credit or charges the customer a higher price for credit based upon a credit report, the company must comply with certain disclosure requirements. There is a growing trend in which companies utilize big data and predictive analytics to make such credit eligibility determinations. FTC REPORT 15–16. Perhaps FCRA can be used to trigger the type of transparency requirements that we propose. Also, in many states, insurance companies that use algorithms are subject to some transparency requirements. *See, e.g.*, Colo. Rev. Stat. Ann. § 10-3-1104.9; Cal. Civ. Code § 1798.145; Conn. Pub. Act. 22-15 ("An Act Concerning Personal Data Privacy and Online Monitoring," effective July 1, 2023).

sex, namely, where height should otherwise be irrelevant (e.g., for the marketing of computer coding classes).[67]

C. A Right to Algorithmic Transparency? Hard Case: Black-Box Algorithms

We now turn to black-box algorithms—machine-learning algorithms, where the process of manipulating inputs to generate outputs is opaque, even to the programmer. With white-box algorithms it is much easier to predict and then confirm a given instance of consumer harm. With black-box algorithms the identification and measurement of harm is more challenging. But it is not impossible. Computer scientists and economists have developed methods to "open" the black-box and "interpret" the algorithm's decision-making process. And, in some cases, the harm caused by the algorithm, i.e., the algorithm's output, can be identified, even without fully understanding how the black-box algorithm generated that harmful output. With black-box algorithms, policymakers need to force the creation of information before they can require its disclosure. This requires an expansion of the right to algorithmic transparency—an expansion that may well be necessary given the growing use of black-box algorithms.[68]

The proposed expansion of the right to algorithmic transparency builds on methods, developed by computer scientists and economists, to interpret black-box algorithms. Regulators could require that companies implement these methods to identify algorithmic harm from price discrimination and quality discrimination, and also from discrimination based on race and sex. The regulator would need to define the transparency-generating methods to be used by firms. Alternatively, firms could be required to disclose their code and their data and the regulator would then implement these methods itself.

To see what we have in mind, begin with the case of discrimination on the basis of race or sex. It is sometimes suggested that if the goal is to ferret out discriminatory motives or to police discriminatory impact, opaque black-box algorithms present special challenges.[69] But in important respects, even black-box algorithms are highly transparent, or at least can be made to be.[70] Certainly they can be far more transparent than human beings, who might not even know their own motivations. In some cases, algorithms can even serve as "discrimination detectors."[71] With the right legal and regulatory systems in place, algorithms can serve as something akin to a Geiger counter that makes it easier to detect—and hence prevent—discrimination. The use of algorithms

[67] Relying on market forces and public pressure is not without risk. For example, as noted above (infra note 50), in some situations accounting for race or sex, or for variables that correlate with race or sex, can help historically disadvantaged groups. And yet public opinion might not reflect a nuanced understanding of when accounting for race or sex is harmful v. helpful.

[68] The growing use of black-box algorithm may be attributed to their greater effectiveness. It may also be attributed to the advantage they offer in terms of avoiding liability. The legal reforms discussed in this section are designed to minimize this advantage.

[69] *See, e.g.*, Nicol Turner Lee, Paul Resnick & Genie Barton, *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*, BROOKINGS (May 22, 2019).

[70] *See* Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Algorithms as discrimination detectors*, 117 PROC. NAT'L ACAD. SCI. 30096, 30096 (2020); Kleinberg et al., *supra* note 50, at 23.

[71] Kleinberg, *Algorithms as discrimination detectors*, *supra* note 90.

can offer far greater clarity and transparency about the ingredients and motivations of decisions.[72] But for them to do that, they must themselves be transparent.

Suppose that algorithms are being asked to solve some prediction problem (say, about who will buy certain products) and that marketing campaigns will be based on those solutions. If algorithms are considering race or gender (by, for example, offering certain products to women but not to men) it should be easy to see that they are doing so—by scrutinizing the algorithm's inputs. If that is what they are doing, they can be rebuilt so as to be blind to any such characteristics.[73] The bigger challenge is when the algorithm considers some factor that is correlated with race or gender. Suppose that the algorithm—deprived of direct data on consumer's gender—learns to use height, because it is correlated with gender and thus can serve as a pretty good proxy for gender. And the challenge becomes even greater when the algorithm finds proxies for consumer ignorance and imperfect rationality, rather than for race or sex, and uses these proxies to discriminate against vulnerable consumers. Below we propose different approaches for meeting this challenge.

The expanded right to algorithmic transparency developed in this Section can serve different policy goals. First, by enhancing transparency about algorithmic harms, it can facilitate public scrutiny and market discipline (as discussed in Sec. B above). If, for example, algorithms are taking advantage of an absence of information or behavioral biases, the public might learn about it—and the practices might stop. There is some reason to believe that a public outcry about relevant practices could change corporate behavior.[74] Second, transparency can serve as a basis for a more heavy-handed regulatory response, when it reveals harm beyond a certain threshold. At the same time, applying the proposed transparency protocols can show that any consumer harm falls below a certain threshold and thus serve as a safe harbor against regulatory scrutiny. Third, a transparency requirement can buttress legal doctrines that require knowledge or intent as a condition for liability. Consider liability for unfair or deceptive acts or practices under Section 5 of the FTC Act. Or consider disparate treatment liability. In certain contexts, the law already requires transparency. For example, when credit is denied, the consumer is entitled to an

[72] *See, e.g.*, Kleinberg, *Algorithms as discrimination detectors*, *supra* note 90.

[73] Although such blinding might end up harming the protected group. *See*, *e.g*, Talia B. Gillis & Jann L Spiess, *Big Data and Discrimination*, 86 U. Chi. L. Rev. 459 (2019); Talia B. Gillis, *The Input Fallacy*, 106 Minn. L. Rev. 1175 (2022).

[74] Lavender Yang, Nicholas Z. Muller & Pierre Jinghong Liang, *The Real Effects of Mandatory CSR Disclosure on Emissions: Evidence from the Greenhouse Gas Reporting Program* 2 (Nat'l Bureau of Econ. Rsch., Working Paper No. 28984, 2021); Archon Fung & Dara O'Rourke, *Reinventing Environmental Regulation from the Grassroots Up: Explaining and Expanding the Success of the Toxics Release Inventory*, 25 ENV. MGMT. 115 (2000). In the policing context, the National Institute of Standards and Technology (NIST) has been testing face recognition algorithms for accuracy. NIST does not formally certify these algorithms. But it issues public reports with vendor-specific performance data, and it publishes on its website a dynamic "leaderboard" ranking algorithm performance. These evaluations and rankings provide incentives for vendors to design better algorithms, as evidenced by vendors' frequent citation to their NIST standings in press and sales materials. *See* Barry Friedman et al., *Policing Police Tech: A Soft Law Solution*, 37 BERKELEY TECH. L.J. (forthcoming 2022); Samuel Dooley et al., *Robustness Disparities in Commercial Face Detection*, OPEN REVIEW (pre-print) (August 2021), https://arxiv.org/pdf/2108.12508.pdf (discussing the role of NIST testing as a "guardrail that has spurred positive, though insufficient, improvement and widespread attention"); Kate Kaye, This little-known facial-recognition accuracy tests has big influence, iapp (Jan. 7, 2019) https://iapp.org/news/a/this-little-known-facial-recognition-accuracy-test-has-big-influence. Friedman et al have called for a formal certification, or pre-approval, requirement for algorithms and other technology used by police forces. *See* Friedman et al., *supra*.

explanation, and many states impose transparency requirements on insurance companies that utilize algorithms[75] The analysis in this Section provides guidance for the implementation of these laws.

We begin, in subsection 1, by describing methods, developed in the computer science and economics literatures, that allow us to peer into the algorithmic black box. We call these methods "transparency protocols." Then, in subsection 2, we explain how policy makers can utilize the transparency protocols to mitigate algorithmic harms. Finally, in subsection 3 we extend standard disparate impact analysis to scrutinize the outcomes of black-box algorithms.

1. Transparency Protocols

While it is impossible (in most cases) to attain a full understanding of how a black-box, machine-learning algorithm operates, computer scientists and economists have developed ways, transparency protocols, that allow us to identify the main decision drivers, i.e., the variables that significantly affect the algorithm's decisions. Here we describe two such protocols.

Teacher-Student. In this protocol, the main black-box algorithm, referred to as the "Teacher" algorithm, trains a simpler, interpretable "Student" algorithm.[76] Specifically, the regulator defines the structure and complexity of the Student. For example, the regulator can specify that the Student will be an easily-interpretable decision-tree algorithm with a depth of three layers. The protocol would then search for the 3-layer tree that most-closely approximates the decisions made by the Teacher algorithm. For example, in the context of algorithmic price discrimination, the regulator could apply the protocol and observe the consumer characteristics that drive pricing decisions in the best Student algorithm (i.e., in the best 3-layer tree).

Linear model. This protocol seeks out a linear model that most-closely approximates the decisions made by the black-box algorithm. Using standard linear-regression techniques, this method searches for a linear combination of consumer characteristics that most-closely predicts the outcomes—e.g., prices, product offers—produced by the black-box algorithm. A challenge with this method is that the resulting linear model would include a very large number of characteristics, limiting the model's interpretability. This challenge is met by utilizing sparsity-creating methods, like LASSO ("least absolute shrinkage and selection operator"), to limit the

[75] The FTC enforces laws that require explainability, e.g., explain why credit was denied or what factors affect your credit score. *See* Andrew Smith, *Using Artificial Intelligence and Algorithms*, Federal Trade Commission, https://www.ftc.gov/business-guidance/blog/2020/04/using-artificial-intelligence-algorithms. *See also* Colo. Rev. Stat. Ann. § 10-3-1104.9; Cal. Civ. Code § 1798.145; Conn. Pub. Act. 22-15 ("An Act Concerning Personal Data Privacy and Online Monitoring," effective July 1, 2023).

[76] See Max Biggs, Wei Sun & Markus Ettl, *Model distillation for revenue optimization: Interpretable personalized pricing*, 139 Proceedings of Machine Learning Research, 946 (Marina Meila & Tong Zhang, eds., 2021) http://proceedings.mlr.press/v139/biggs21a/biggs21a.pdf (developing a method of translating a complex non-parametric prediction model into a simple pricing policy based on a decision tree. The leaves contain (user, item) pairs with similar optimal prices.) Follow up works include: Shivaram Subramanian, Wei Sun, Youssef Drissi & Markus Ettl, *Constrained prescriptive trees via column generation*, Proceedings of the 36th AAAI Conference on Artificial Intelligence (2022) (allowing constraints, such as (i) requiring that all consumers are charged the same price except for loyalty-card holders who are charged a lower price; or (ii) requiring that one item (say an economy ticket) is priced at least X dollars less then another (a business class ticket).) The general approach, called "knowledge distillation," was developed by Geoffrey Hinton, Oriol Vinyals, & Jeff Dean, *Distilling the knowledge in a neural network*, arXiv preprint arXiv:1503.02531 (2015).

number of characteristics, such that the linear model includes only those characteristics that have the largest effect on the outcome.[77]

2. Applying the Transparency Protocols

There are (at least) two possible approaches for applying the transparency protocols, depending on whether the regulator has access to an identifiable "protected characteristic."

Without an identifiable "protected characteristic": Look for suspicious characteristics. The idea behind this approach is straightforward: Apply a transparency protocol to identify the consumer characteristics that exert significant influence over the algorithm's decision-making process, and target scrutiny—market scrutiny and regulatory scrutiny—towards "suspicious characteristics." For example, the regulator might observe that height plays an important role in the decision-making process—shorter consumers are offered higher prices, perhaps because height is correlated with gender. Or the regulator might observe that consumers with little savings and a lot of debt are offered treadmills or gym subscriptions, perhaps because limited savings and significant debt are correlated with present bias.[78] The role played by such seemingly-irrelevant, suspicious characteristic could trigger regulatory scrutiny or it could be made public and trigger a market reaction.

We recognize, of course, that what counts as a *suspicious* variable might not be obvious. Still, there will be cases, as in the examples we offered, where it is clear that the weight placed, by the algorithm, on a consumer characteristic can be explained only by that characteristic's correlation with the consumer's race, sex or bias.[79] Moreover, any concern about the identification of the suspicious-characteristic criterion should be mitigated to the extent that the transparency exercise is designed to trigger a market reaction. Then the market, rather than a regulator, would decide whether the firm has a convincing reason to set higher prices for shorter consumers, for instance.

[77] CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING: A GUIDE FOR MAKING BLACK BOX MODELS EXPLAINABLE, Sec. 5.1. (Independently published, 2022), https://christophm.github.io/interpretable-ml-book/. There are other interpretability methods, in addition to the two methods described in the text. *See*, *e.g.*, Laura Blattner & Jann Spiess, *Machine Learning Explainability and Fairness: Insights from Consumer Lending*, FinRegLab Empirical White Paper, April 2022; Laura Blattner, Scott Nelson & Jann Spiess, *Unpacking the Black Box: Regulating Algorithmic Decisions*, working paper (2021). [Perhaps discuss additional interpretability methods in the text.] The different transparency protocols, or interpretability methods, are not without limits. In particular, different protocols can yield different sets of important, decision-affecting variables. Indeed, even when utilizing a single transparency protocol, we may get different sets of important, decision-affecting variables. For example, there can be several 3-layer trees that approximate the decisions of the Teacher algorithm at more or less the same degree of precision. One response to these limitations is to use multiple transparency protocols and, with each protocol, to consider multiple outcomes (i.e., multiple sets of important, decision-affecting variables).

[78] As noted above, it is not clear that a high level of borrowing and a low level of savings necessarily implies myopia or present bias; it could also imply exponential discounting with a high discount rate, i.e., it could imply a (rational) preference rather than a bias. But a rational consumer with such a high discount rate would not get a gym subscription.

[79] There is a question of whether the regulator should announce, in advance, what factors would be considered suspicious. In any event, over time firms will learn what characteristics are more likely to trigger scrutiny. If firms know that a variable would trigger scrutiny, they may exclude this variables from the data that is fed into the algorithm. The algorithm would then find another variable that is correlated with the excluded variable. This other variable would likely be equally suspicious. The transparency approach may thus lead to the gradual removal of variables that are likely to trigger consumer harm.

Does an identifiable "protected characteristic" emerge as a key decision driver? With this approach, the regulator would again apply a transparency protocol to identify the consumer characteristics that exert significant influence over the algorithm's decision-making process. But now the question is not whether one of the influential characteristics is suspicious. Rather the question is whether one of the influential characteristics is a previously identified "protected characteristic." If the protected characteristic emerges as a key decision driver, then regulatory or market scrutiny should follow.[80]

It is straightforward to identify race or sex as a protected characteristic and see whether they emerge as influential decision drivers when applying a transparency protocol. But, as explained above, the main concern is about a different protected characteristic—the consumer's information or rationality deficit. And it is more challenging to identify a protected characteristic variable that distinguishes between informed and uninformed consumers or between biased and unbiased consumers. Can this category of less-sophisticated, imperfectly rational consumers be identified in advance? We suggest that, at least in some cases, the answer is 'yes.'

Specifically, biases, misperceptions and other deviations from perfect rationality can be measured using survey evidence. For example, in the health insurance context, Baillon et al used survey evidence to measure (i) overestimation of risk (of incurring medical expenses) and (ii) the shape of the consumer's Prospect Theory utility function.[81] And, in the consumer credit context, Meier and Sprenger used incentivized choice experiments to measure subjects' level of present bias.[82] Assuming that generally administered surveys (like the Survey of Consumer Finances) provide data on a sufficiently large subset of a seller's customer base, and that such surveys could be amended to include bias-measuring questions, these surveys can be used to define, in advance, a protected class of biased or imperfectly rational consumers. Alternatively, regulators may be able to use measures of, or proxies for, sophistication, such as the level of education or experience in the relevant context,[83] and then treat limited sophistication as a protected characteristic.

In the previous approach, we did not have a protected-characteristic variable and so the search for influential characteristics could only produce suspicious variables that correlate with the protected characteristic. Now we have a protected-characteristic variable and the question is whether the search for influential characteristics would identify this variable as influential. The "suspicious characteristics" approach has an important advantage: The regulator is not required to define, in advance, a protected characteristic. The crux of that approach was identifying the

[80] In implementing this approach, we should be cognizant of possible correlations between the protected-characteristic variable and other variables. For example, if present bias is highly correlated with limited retirement savings, then the interpretable model would select either the present-bias variable or the retirement savings variable. A possible response to this concern is to consider not one, but several interpretable models, i.e., to consider not only the best interpretable model (as defined by some transparency protocol), but perhaps the top three models.

[81] Aurélien Baillon et al. 2022. A behavioral decomposition of willingness to pay for health insurance. Journal of Risk and Uncertainty 64: 43-87. Baillon et al show that while overestimation of risk pushes the WTP up, the risk-loving feature of the PT utility function pushes it down, with the latter effect dominating.

[82] Meier, Stephan, and Charles Sprenger. 2010. "Present-Biased Preferences and Credit Card Borrowing." *American Economic Journal: Applied Economics*, 2 (1): 193-210. Meier and Sprenger showed that biased consumers buy different products and use these products differently than unbiased consumers. We realize that incentivized choice experiments are more than simple surveys, such that the relevant evidence can be more difficult to obtain.

[83] Compare: the sophisticated investor test in the securities context. *See, e.g.*, *Terra Sec. ASA Konkursbo v. Citigroup, Inc.*, 820 F. Supp. 2d 541, 545–46 (S.D.N.Y. 2011).

characteristics that exerted the most influence on the algorithm's decision-making process, relying on ex post "suspiciousness" scrutiny by the regulator or by the market. When we have an identifiable "protected characteristic," we can avoid the "suspiciousness" criterion but, of course, the regulator must be able to define, in advance, what the protected characteristic is and there must be an objective way to identify or measure this protected characteristic.

3. Disparate Impact on Consumers with an Identifiable "Protected Characteristic"

As with the "protected characteristic as a key decision driver" approach, here too the regulator must specify, in advance, what the protected characteristic is. In the spirit of the disparate impact doctrine, this approach evaluates the algorithm's decisions, or outcomes, and targets scrutiny towards cases where consumers with a protected characteristic are treated differently. This approach has been developed in the context of discrimination based on race or sex as protected characteristics, and we propose to extend it to discrimination based on imperfect information or imperfect rationality as protected characteristics.

As with any disparate impact analysis, the challenge is that consumers with a protected characteristic may be treated differently, because the protected characteristic is correlated with other, relevant (and not protected) characteristics. For example, imperfect information or imperfect rationality may be correlated with income or preferences. In the case of discrimination based on race or sex, the doctrinal question is whether "similarly situated" consumers were treated differently.[84] The same question should be asked when the protected characteristic is bias or misperception: whether biased consumers were treated differently from "similarly situated" unbiased consumers.

Regulators can address this challenge by using a linear regression model to evaluate how different consumer characteristics affect the algorithm's decisions. The model would include the protected-characteristic variable, say a measure of present bias, and the coefficient assigned to that variable would measure the effect of present bias on the outcome. The model will also include other relevant (not protected) characteristics, like income. By including these other, control variables, regulators can compare between "similarly situated" consumers. In our example, the coefficient assigned to the present-bias variable would measure the effect of the bias on the outcome for consumers with the same income level. If this effect is significant, then regulatory or market scrutiny should follow.

How do we select the set of control variables? Put differently, how do we define what counts as "similarly situated"? Should we include only income? Should we add the consumer's wealth? Credit rating? Zip code? The appropriate control variables are context dependent.[85] The regulator can use its subject-matter expertise to select these variables. Or we can use sparsity-

[84] *See* Gillis & Spiess, *supra* note 73 (arguing for disparate-impact-type analysis of outcomes and noting the challenge of defining "similarly situated" consumers). *See also* Gillis, *supra* note 73.

[85] In the consumer credit context these variables will include standard underwriting variables, such as FICO score, loan-to-value ratio, debt-to-income ratio, loan amount, type of loan, etc'. *See* Ian Ayres, Gary Klein & Jeffrey West, *The Rise and (Potential) Fall of Disparate Impact Lending Litigation*, in Lee Anne Fennell and Benjamin J. Keys, eds, *Evidence and Innovation in Housing Law and Policy* 231, 236 (Cambridge 2017) (analyzing *In re Wells Fargo Mortgage Lending Discrimination Litigation*, 2011 WL 8960474 (ND Cal), in which plaintiffs used regression analysis—including models with fewer controls and models with many controls—to prove unjustified disparate impacts).

creating methods, like LASSO, to select the control variables, or consumer characteristics, with the largest effect on the outcome.[86] Note that the way we propose to use the linear model in is different from the way it was used in the "suspicious characteristics" approach or in the "protected characteristic as a key decision driver" approach. In the "suspicious characteristics" approach, regulators did not have a measurable protected-characteristic variable, and the goal was to identify characteristics that have a large effect on the algorithm's decisions and scrutinize the suspicious ones. In the "protected characteristic as a key decision driver" approach, regulators had a measurable, protected-characteristic variable and the goal was to see if this variable has a large effect on the algorithm's decisions. Here, regulators have a measurable protected-characteristic variable, and the goal is to assess the effect of this variable on "similarly situated" consumers (where "similarly situated" is defined by the control variables).

A related approach would assess the disparate impact of the algorithm, relative to the pre-algorithm baseline. To implement this approach, the regulator would need data on outcome decisions, e.g., pricing decisions, before and after the black-box, pricing algorithm was adopted. The regulator would then run the regression model, with the same explanatory variables—the same protected-characteristic variable and the same control variables—on pre- and post-algorithm outcome data. If the coefficient assigned to the protected-characteristic variable is larger when the algorithm sets prices, then the disparate impact on the protected group was made worse by the algorithm.

D. Regulating the Design and Implementation of Algorithms

Algorithms come in many shapes and sizes, and some are more harmful than others. If the harms are sufficiently severe, regulators might intervene in the design process. In addition, courts can police especially harmful algorithms under a model of liability for defective products. In some cases, regulators could impose requirements on the data that are used to train machine learning algorithms. An obvious, simple-minded requirement is 'exclude information about race' (simple-minded, because race can be accurately inferred from correlates and excluding a direct race variable could, in some cases, actually harm racial minorities). But other, more sensible requirements might be imposed on the training data, such as requiring balanced representation of different groups of consumers and excluding biased data.[87]

In addition, regulators might require that algorithms be programmed with certain constraints. For example, computer scientists and others have explored different mathematical formulations of fairness or equality constraints that can be imposed on the algorithm.[88] Specifically, Cohen et al. propose four definitions of "fairness," with the most relevant being "price fairness," i.e. that "prices offered to the two groups are nearly equal."[89] To date, this work has generally focused on race and sex, requiring that men and women be offered nearly equal prices,

[86] If a control variable is closely correlated with the protected-characteristic variable, then we might run into a multi-collinearity problem.

[87] *See also* FTC REPORT 27–28 (discussing the importance of ensuring representation and elimination of biases in data sets).

[88] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Ashesh Rambachan, *Algorithmic Fairness*, 108 AM. ECON. REV. PAPERS & PROCS. 22, 22 (2018).

[89] Maxime C. Cohen, Adam N. Elmachtoub & Xiao Lei, *Price discrimination with fairness constraints*, Management Science (2022), https://maxccohen.github.io/Pricing-Fairness.pdf.

or that whites and Blacks be offered nearly equal prices. But it could be applied to consumer bias or misperception, if they can be defined and measured (as explained in Section C.2. above). Regulators could then require that biased and unbiased consumers are charged (nearly) the same prices or offered the same products.[90, 91]

At the extreme, would it be desirable to prohibit certain uses of algorithms? For reasons we have sketched, there is no sufficient justification for doing so in PI-PR cases, except perhaps if sex-based on race-based discrimination is identified. But in II-IR cases, there is a real question whether it might be appropriate to forbid the use of algorithms to make distinctions with respect to prices and product characteristics. In principle, such a prohibition could benefit consumers in the circumstances we have discussed. If regulators could devise a fine-tailored intervention, and apply it only in those circumstances, they would by hypothesis increase consumer welfare.[92]

Such prohibitions could be viewed as the continuation, in the algorithmic context, of behaviorally informed policies forbidding practices that exploit behavioral biases. Consider the CARD Act, enacted in 2010, which imposes regulatory restrictions on late fees and overuse fees. Those restrictions are best understood as an effort to respond to II-IR situations, which have been particularly pronounced among people with poor credit ratings.[93] The central idea is that fees of this kind are not transparent to consumers and that, for credit card companies, they operate essentially as rents.[94] In these circumstances, regulatory restrictions—in this case in the form of price caps—could be taken as a response to a kind of behavioral market failure, and they should be effective if companies are not, in fact, competing over the relevant product characteristics. Indeed, the evidence suggests that consumers have gained almost $12 billion annually as a result of the restrictions, with particular benefits for people who are struggling economically.[95] To the extent that algorithmic harm is being imposed in II-IR situations, the argument for restrictions of that kind gains force.

To be sure, there are serious problems of administrability. Regulators do not, of course, deal with binary cases of PI-PR and II-IR. They deal with heterogenous populations, with complex mixes of information and rationality. If regulators were themselves perfectly informed, they would be able to make a judgment about the net benefits of any ban. They would be able to identify the

[90] As discussed in Section C.3. above, it may be justified to set higher prices for consumers with protected characteristics, if these characteristics are correlated with other, relevant (and not protected) characteristics. For example, race may be correlated with income, or gender may be correlated with preferences. Therefore, the fairness constraint needs to be defined as: "similarly situated" consumers must be treated similarly, where "similarly situated" is operationalized as discussed in Section C.3.

[91] The FTC has warned firms that use algorithms to avoid disparate impact. *See* Smith, *supra* note 96 ("You can save yourself a lot of problems by rigorously testing your algorithm, both before you use it and periodically afterwards, to make sure it doesn't create a disparate impact on a protected class.") A standard defense against a disparate-impact antidiscrimination claim is "business necessity." Computer scientists have quantified the cost, in terms of lost-profits, of imposing different fairness constraints on the algorithm. Such analysis should inform any assessment of "business necessity." Specifically, if a non-discrimination constraint reduces profits by a relatively small amount, then the "business necessity" defense should be rejected.

[92] *Cf.* Paterson et al., *supra* note 38, at 12–13, 14–16 (2021) (discussing bans, and considering the use of something like the FTC's § 5 powers to police algorithms).

[93] *See* Natasha Sarin, *Making Consumer Finance Work*, 119 COLUM. L. REV. 1519, 1524–25 (2019).

[94] *See id.*

[95] *See* Simut Agarwal, Souphala Chomsisengphet, Neale Mahoney & Johannes Stroebel, *Regulating Consumer Financial Products: Evidence from Credit Cards*, 130 Q. J. ECON. 111, 113 (2015).

circumstances in which algorithms would, on balance, do more harm than good (and perhaps hurt people at the bottom of the economic ladder[96]). Lacking perfect information, they might do best to keep prohibitions in the toolkit, but reserve them for the most obvious or egregious cases.

E. Applying the Reforms to the Different Harm Categories

In developing legal reforms, the preceding discussion mentioned examples of algorithmic harms that the reforms were designed to address. We now explore how the proposed reforms can be applied to address the main categories of algorithmic harm analyzed in earlier parts of the Article.

Algorithmic Price Discrimination. One of our main concerns throughout has been algorithmic pricing that targets consumers' biases and misperceptions. In implementing our proposed reforms, a main challenge involves identifying instances of such targeting, especially when sellers employ black-box pricing algorithms. Our discussion of algorithmic transparency suggested several ways for meeting this challenge. First, regulators can use, or force sellers to use, transparency protocols—to identify variables that exert significant influence over the algorithm's pricing decisions. If any of these variables is "suspicious," i.e., its influence can be explained only as a proxy for consumer bias or misperception, then regulatory or market scrutiny should follow. For example, if the pricing algorithm used by a credit card issuer places significant weight on the consumer's retirement savings, this may be considered suspicious—especially if low savings trigger offers with low introductory interest rates and high long-term rates, perhaps because the algorithm associates low savings with present bias.

Second, if a specific bias or misperception can be measured, e.g., through generally-administered surveys, regulators could use transparency protocols and see if the measured bias or misperception emerges as one of the key decision drivers. Under the disparate impact approach, which also applies when a specific bias or misperception can be measured, regulators can test for special harm on those who suffer from such a bias or misperception: are consumers with high bias levels charged higher prices than consumers with low bias levels who are otherwise "similarly situated"? Finally, if the transparency regulations reveal bias-based price discrimination, this could potentially trigger liability, e.g., under Section 5 of the Federal Trade Commission Act, which prohibits unfair practices, or similar state UDAP statutes. In the consumer credit context, where the prohibition extends also to abusive practices, it would be even easier to impose liability. And, as mentioned above, the transparency reforms, which force an opening of the algorithmic black box, would prevent sellers from claiming that they did not know that their algorithms were discriminating. To be clear: It is not our purpose here to conclusively identify a specific doctrinal source of liability. This would require an analysis of doctrinal and policy considerations for and against using a specific doctrine to police algorithmic harms. We relegate such analysis to future work. The conceptual point is that algorithmic transparency can help make the case for imposing liability.

Beyond transparency, policymakers can regulate the design and implementation of pricing algorithms to reduce the risk of bias-based pricing. Here too, black-box algorithms pose a

[96] *See generally* Matthew D. Adler, *Theory of Prioritarianism*, *in* PRIORITARIANISM IN PRACTICE (Matthew D. Adler ed., forthcoming 2022) (outlining the theory of prioritarianism as a branch of welfare consequentialism).

challenge and the proposed solution requires that the specific bias or misperception be measurable. If such measurement is possible, then regulators can force sellers and algorithm designers to include a no-discrimination constraint in the algorithm's code—to ensure that biased consumers are charged (nearly) the same price as non-biased consumers.[97]

Algorithmic Quality Discrimination. While this category of harm is distinct from the previous category, the proposed legal reforms apply in a similar way. The main difference is that regulators now need to ask what affects the algorithm's product-targeting decisions, rather than pricing decisions, and, relatedly, whether biased consumers are offered inferior products. For example, it would be suspicious if a consumer's low rate of retirement savings significantly influences the algorithm's decision to offer the consumer a gas guzzler rather than a hybrid car. And if the level of consumers' present bias can be measured, regulators would want to know if biased consumers are more likely to be offered gas guzzlers, as compared to "similarly situated" unbiased consumers. Finally, if the transparency regulations reveal bias-based quality discrimination, this could trigger liability for unfair or abusive practices. Moving beyond transparency-related reforms, regulators can force sellers and algorithm designers to include a no-discrimination constraint in the algorithm's code—to ensure that a consumer's bias level does not affect the type of car that this consumer is offered.

Algorithmic Discrimination Based on Race and Sex. The focus on race and sex makes the regulator's job easier. As explained above, many of the proposed reforms can be applied only if the protected characteristic is identifiable or measurable. This condition is more easily met with discrimination that is based on race and sex, as many data sets that are used by algorithms in consumer markets include information on the consumers' race and sex. Therefore, it would be easier to identify disparate impact—in terms of pricing or product targeting—on women and minorities, for example. It would be similarly easier to know whether race or sex exerted significant influence in the algorithm's decision-making process (suggesting the presence of disparate treatment). And it would be easier to force sellers and algorithm designers to include a no-discrimination constraint in the algorithm's code.

The proposed reforms could be applied even if the algorithms are denied access to information about the consumer's race or sex, perhaps in attempt to comply with antidiscrimination laws. As a preliminary matter, and as we explained above, removing this information from the data is unlikely to prevent discrimination, as the algorithm would likely find other variables that correlate with race or sex. In terms of the proposed reforms, the regulator can require submission of the full data, including the race and sex variables, if the regulator wants to apply the transparency protocols itself. Or it could force the seller to apply the transparency protocols using the full data. And, if the regulator decides to go beyond transparency and force

[97] With respect to algorithmic behavior-based price discrimination, we note that discrimination based on past purchasing behavior should be relatively easy to detect, in the sense that a previous decision by the consumer—to buy or not to buy the product at an offered price—is identifiable and measurable. Therefore, the transparency reforms that we have outlined should be relatively easy to implement: it should be relatively easy to learn, and to inform the market, that a seller's algorithm sets different prices to consumers based on their past purchasing behavior. Recall that behavior-based pricing is harmful especially when consumers are not aware of this pricing strategy. By informing consumers about the seller's pricing strategy, the transparency requirement directly targets, and potentially eliminates, a precondition for consumer harm. Indeed, as noted in Sec. III.C., when consumers know about the BBP (and react strategically), sellers lose from BBP and would like to commit not to utilize BBP. The transparency reforms would facilitate such a commitment.

sellers and algorithm designers to include a no-discrimination constraint in the algorithm's code, the mandate would require that the constraint be implemented using the race and sex variables.

VII. Conclusion

Machine learning algorithms are increasingly able to predict what goods and services particular people will buy, and at what price. In many cases, the use of algorithms promises to increase efficiency and to promote social welfare; it might also promote fair distribution. But when consumers suffer from an absence of information or from behavioral biases, algorithms can cause serious harm. Behaviorally informed disclosure requirements would reduce the risks that algorithms might exploit ignorance or bias, and to that extent, the argument for those requirements is increasingly strong. Transparency about the nature, uses, and consequences of algorithms would also be a relatively modest and potentially effective response. In appropriate cases, regulators can police the design and implementation of algorithms, imposing constraints that range from mild to stringent. More general legal bans on exploitation, by algorithms, of imperfect information and behavioral biases would be an excellent idea in principle, but would create serious problems of administrability. In the fullness of time, regulators are likely to find ways to overcome those problems.

We recognize that it will be challenging for lawmakers to intervene in a welfare-enhancing way, as the harmful uses of algorithms would need to be distinguished from the beneficial ones. We thus address most of the proposed legal reforms to regulators with market-specific expertise and encourage these regulators to continue to develop algorithmic expertise. With respect to the domain of analysis, while we have focused on algorithmic harms in consumer markets, similar harms arise in labor markets, and the legal reforms that we have suggested can also apply, with appropriate adjustments, in the labor context.

Appendix

The Appendix provides formal models of (i) the Behavior-Based Pricing (BBP) analysis from Sec. III.C, and (ii) the Algorithmic Quality Discrimination analysis in II-IR markets from Sec. IV.B.

I. Behavior-Based Pricing

A. II-IR Markets

With BBP, it is easier to start with a version of the II-IR case, namely, the case of uninformed consumers who are not aware of the seller's BBP. These consumers will not adjust their early-period purchasing decisions to secure lower later-period prices. To ascertain the effect of algorithmic BBP, we begin with the pre-algorithmic benchmark. In this pre-algorithmic world, a monopolistic seller will set the same (monopoly) uniform price in both the early and late periods. With algorithmic BBP, the seller will set a uniform, higher early-period price and two late-period prices—a higher price for consumers who purchased in the early period and a lower price for those who did not. The lower late-period price allows poorer, lower-WTP consumers who did not purchase in the early period to enter the market. The higher late-period price extracts more surplus from the richer, higher-WTP consumers who made an early-period purchased.

The overall welfare effects of BBP are nuanced. From an efficiency perspective, with BBP sellers serve more consumers in the later period (thanks to the differentiated pricing), but fewer consumers in the early period (because of the higher early-period price). From a distributional perspective, higher-WTP consumers who are likely richer are harmed by the higher prices in both the early and late periods. At the same time, some lower-WTP consumers, who are likely poorer and were excluded from the market without BBP, are able to participate in the market and gain surplus with BBP. It will often be the case that consumers as a group are harmed by BBP, whereas a subgroup of poorer consumers benefits. The overall welfare assessment of algorithmic BBP is complicated by these tradeoffs. To illustrate the effects of BBP and gain further insight into the tradeoffs that determine the normative evaluation of this practice, we next study a detailed example of BBP.

Setup. Consider a product that gives each consumer a value $v \in [0, V]$, and let the probability density function $f(v)$, and the corresponding cumulative distribution function $F(v)$, represent the distribution of values across a unit mass of consumers. For simplicity, we assume a uniform distribution, such that $f(v) = \frac{1}{V}$ and $F(v) = \frac{v}{V}$.[1] The distribution of values determines the demand for this product: For any price p, the quantity sold is given by $q(p) = 1 - F(p)$, i.e., consumers with a value that exceeds the price will purchase the product. At this price p, the monopolistic seller makes a profit of $\pi(p) = p \cdot q(p)$, if we normalize the per-unit cost of production to zero; and the consumer surplus is: $\int_p^V (v - p) f(v) dv$, aggregating the net benefit, $v - p$, across consumers with values $v \in [p, V]$ who purchase the product at the price p. There are

[1] For example, half of all consumers get a value of $\frac{V}{2}$ or less from the product, i.e., $F\left(\frac{V}{2}\right) = \frac{V/2}{V} = \frac{1}{2}$.

two time periods, period 1 and period 2. We assume that, in each period, each consumer purchases one unit of the product, at most. For simplicity, we assume no time discounting.

Pre-algorithmic world. In the pre-algorithmic world, the monopolist cannot engage in BBP. Therefore, it will set the same price in both periods, and this price will be offered to all consumers. Specifically, the offered price will be the standard monopoly price, which is $p^S = \frac{V}{2}$ in our setup.[2] Accordingly, consumers with above-median values purchase the good, whereas consumers with below-median values are excluded from the market. The monopolist's profit is: $\pi(p^S) = p^S \cdot q(p^S) = \frac{1}{4}V$ in each period, for a total profit of $\frac{1}{2}V$. And the consumer surplus is: $CS(p^S) = \int_{p^S}^{V}(v - p^S)f(v)dv = \frac{1}{8}V$ in each period, for a total consumer surplus of $\frac{1}{4}V$.

Post-algorithmic world. In the post-algorithmic world, the monopolist engages in BBP. It will set a period 1 price p_1, such that high-value consumers, with $v \in [p_1, V]$, buy the product in period 1; and low-value consumers, with $v \in [0, p_1]$, do not buy the product in period 1. The monopolist will then set two different period 2 prices—one price p_2^H for the high-value consumers who bought the product in period 1, and another, lower price p_2^L for the low-value consumers who did not buy the product in period 1. In period 1, the seller is facing the entire market, and demand is given by $q_1(p_1) = 1 - F(p_1)$. The seller's profit is: $\pi_1(p_1) = p_1 \cdot q_1(p_1)$; and the consumer surplus is: $CS_1(p_1) = \int_{p_1}^{V}(v - p_1)f(v)dv$.

In period 2, for the high-value segment, covering all consumers with $v \in [p_1, V]$, demand is given by $q_2^H(p_2^H) = 1 - F(p_2^H)$.[3] The seller's profit is: $\pi_2^H(p_2^H) = p_2^H \cdot q_2^H(p_2^H)$; and the consumer surplus is: $CS_2^H(p_2^H) = \int_{p_2^H}^{V}(v - p_2^H)f(v)dv$. In our setup, the profit-maximizing price is $p_2^H = p_1$.[4] All of the high-value consumers, with $v \in [p_1, V]$, who purchases the product in period 1 will also purchase the product in period 2. Therefore, we can rewrite the monopolist's profit as: $\pi_2^H(p_1) = p_1 \cdot q_2^H(p_1)$; and the consumer surplus as: $CS_2^H(p_1) = \int_{p_1}^{V}(v - p_1)f(v)dv$. In the low-value segment, covering all consumers with $v \in [0, p_1]$, demand is given by $q_2^L(p_2^L) = F(p_1) - F(p_2^L)$. The seller's profit is: $\pi_2^L(p_2^L) = p_2^L \cdot q_2^L(p_2^L)$; and the consumer surplus is: $CS_2^L(p_2^L) = \int_{p_2^L}^{p_1}(v - p_2^L)f(v)dv$. In our setup, the profit-maximizing price is $p_2^L = \frac{p_1}{2}$.[5] Of the low-value consumers who did not buy in period 1, the upper-half, i.e., consumers with $v \in \left[\frac{p_1}{2}, p_1\right]$ buy the product in period 2. Therefore, we can rewrite the monopolist's profit as: $\pi_2^L\left(\frac{p_1}{2}\right) = \frac{p_1}{2} \cdot q_2^L\left(\frac{p_1}{2}\right)$; and the consumer surplus as: $CS_2^L(p_1) = \int_{p_1/2}^{p_1}\left(v - \frac{p_1}{2}\right)f(v)dv$.

We can now derive the period 1 price. The seller sets this price to maximize the sum of its period 1 profit, $\pi_1(p_1)$, together with the two period 2 profits—$\pi_2^H(p_1)$ for the high-value segment

[2] The monopolist sets a price that solves: $\max_p \pi(p)$.

[3] As long as $p_2^H \geq p_1$. This condition is satisfied (as we show below).

[4] The price that maximizes $\pi_2^H(p_2^H)$ in an unrestricted domain is $\frac{V}{2}$. Since $p_1 > \frac{V}{2}$ (as we show below) and the domain of the high-value segment is $v \in [p_1, V]$, we have a corner solution: $p_2^H = p_1$.

[5] This is the price that maximizes $\pi_2^L(p_2^L)$.

and $\pi_2^L\left(\frac{p_1}{2}\right)$ for the low-value segment.[6] In our setup, the profit-maximizing price is $p_1 = \frac{4V}{7}$, such that the upper-$\frac{3}{7}$ of consumers, with values $v \in \left[\frac{4V}{7}, V\right]$, buy the good in period 1. Then, in period 2, the seller will set $p_2^H = p_1 = \frac{4V}{7}$ for the consumers who bought the product in period 1, such that the same consumers, with values $v \in \left[\frac{4V}{7}, V\right]$, buy also in period 2; and the seller will set $p_2^L = \frac{p_1}{2} = \frac{2V}{7}$ for the consumers who did not buy the product in period 1, such that consumers with values $v \in \left[\frac{2V}{7}, \frac{4V}{7}\right]$, buy in period 2.

Comparison. BBP clearly increases the seller's profit; otherwise, the seller would avoid BBP and set prices as in the pre-algorithmic world. Specifically, whereas seller's profit was $\frac{1}{2}V$ without BBP, it is: $\pi_1(p_1) + \pi_2^H(p_1) + \pi_2^L\left(\frac{p_1}{2}\right) = \frac{28}{49}V$ with BBP. But, while the seller benefits from BBP, consumers are harmed. Without BBP, consumer surplus was $\frac{1}{4}V$. With BBP, consumer surplus is: $CS_1(p_1) + CS_2^H(p_1) + CS_2^L\left(\frac{p_1}{2}\right) = \frac{11}{49}V$. In our setup, the harm to consumers from BBP, i.e., the reduction in consumer surplus ($\frac{1}{4}V - \frac{11}{49}V$), is smaller than the benefit to the seller, i.e., the increase in the seller's profit ($\frac{28}{49}V - \frac{1}{2}V$), such that BBP increases overall efficiency.[7] Yet, given the adverse distributional effect, BBP may still be socially undesirable.

Drilling down further, we can distinguish between four groups of consumers, as shown in Table 2 below. The table also presents, for each group, the consumer surplus, the seller's profit and the total surplus (which combines the consumer surplus and the seller's profit), with and without BBP.

Consumers with	Consumer Surplus		Seller's Profit		Total	
	No BBP	BBP	No BBP	BBP	No BBP	BBP
$v \in \left[\frac{4V}{7}, V\right]$	$\frac{96}{392}V$	$\frac{72}{392}V$	$\frac{168}{392}V$	$\frac{192}{392}V$	$\frac{264}{392}V$	$\frac{264}{392}V$
$v \in \left[\frac{V}{2}, \frac{4V}{7}\right]$	$\frac{2}{392}V$	$\frac{7}{392}V$	$\frac{28}{392}V$	$\frac{8}{392}V$	$\frac{30}{392}V$	$\frac{15}{392}V$
$v \in \left[\frac{2V}{7}, \frac{V}{2}\right]$	0	$\frac{9}{392}V$	0	$\frac{24}{392}V$	0	$\frac{33}{392}V$

[6] The seller sets a price that solves: $\max_{p_1}\left\{\pi_1(p_1) + \pi_2^H(p_1) + \pi_2^L\left(\frac{p_1}{2}\right)\right\}$.

[7] The result that BBP increases overall efficiency depends on the uniform distribution assumption.

$v \in \left[0, \frac{2V}{7}\right]$	0	0	0	0	0	0

Table 2: Disaggregated Effects of BBP in II-IR Markets

We can now summarize the effect of BBP on each group: (1) Consumers with $v \in \left[0, \frac{2V}{7}\right]$ would be excluded from the market with and without BBP. (2) Consumers with $v \in \left[\frac{2V}{7}, \frac{V}{2}\right]$ would be excluded without BBP and served, albeit only in the second period, with BBP. (3) Consumers with $v \in \left[\frac{V}{2}, \frac{4V}{7}\right]$ would be served in both periods without BBP and only in the second period with BBP. Still, because of the lower price charged with BBP in the second period, they enjoy a higher consumer surplus; and the seller's profit from serving these consumers is lower. (4) Consumers with $v \in \left[\frac{4V}{7}, V\right]$ would be served, in both periods, with and without BBP. BBP allows the seller to charge a higher price, thus shifting surplus from consumers to the seller (the total surplus is not changed by the introduction of BBP). While BBP harms consumers as a group, the distributional effects are more subtle: consumers with $v \in \left[\frac{4V}{7}, V\right]$ who are likely richer are harmed by BBP, whereas consumers with $v \in \left[\frac{V}{2}, \frac{4V}{7}\right]$ and with $v \in \left[\frac{2V}{7}, \frac{V}{2}\right]$ who are likely poorer benefit from BBP.[8] (Consumers as a group are harmed because the group with $v \in \left[\frac{4V}{7}, V\right]$ is larger.)

B. PI-PR Markets

We next turn to PI-PR markets, where consumers are aware of the seller's BBP. As noted above, some high-WTP consumers will strategically refrain from making an early-period purchase in order to secure a lower price in the later period. This reduces efficiency and consumer surplus in the early period. In the later period, the algorithm segments the market, with a higher price for consumers who purchased in the early period and a lower price for those who did not. (From a distributional perspective, the outcome in PI-PR markets is somewhat less attractive, as the lower, later-period price is enjoyed by some relatively wealthy consumers who strategically refrained from purchasing in the early period.) When consumers are aware of the seller's use of BBP and respond strategically, BBP helps consumers and harms sellers. Therefore, in the early period, sellers would prefer to commit to refrain from using BBP, if they could. But such a commitment may well prove impossible: in the later period, armed with reams of data and the algorithms to analyze it, sellers will have a strong incentive to engage in BBP; and sophisticated consumers will anticipate this in the early period and respond accordingly. From a social welfare perspective, algorithmic BBP can be desirable in PI-PR markets.

Post-algorithmic world. Whereas in the II-IR case, in period 1, consumers bought the product whenever the value that they gained from the product exceeded its price, in the PI-PR case

[8] If richer consumers are less likely to be unaware of the seller's BBP strategy and thus less likely to be harmed by BBP (see Section B below), then we should be less concerned about BBP.

consumers might refrain from making a period 1 purchase even if value exceeds price. Therefore, we need to derive a value threshold, $\tilde{v}_1$, such that only consumers with $v \in [\tilde{v}_1, V]$ will buy the product in period 1 (note that $\tilde{v}_1$ will exceed the period 1 price, p_1). At this threshold, the loss from forgoing a beneficial, period 1 purchase exactly equals the gain from a lower, period 2 price: $\tilde{v}_1 - p_1 = p_2^H - p_2^L$; we call this the "threshold equation." The period 2 prices also need to be adjusted, relative to the II-IR case, such that the threshold $\tilde{v}_1$ replaces p_1. Specifically, we have $p_2^H = \tilde{v}_1$ and $p_2^L = \frac{\tilde{v}_1}{2}$. Plugging these period 2 prices into the threshold equation, we get: $\tilde{v}_1(p_1) = 2p_1$. We can also rewrite the period 2 prices as a function of p_1: $p_2^H(p_1) = 2p_1$ and $p_2^L(p_1) = p_1$.

The seller sets p_1 to maximize the sum of its period 1 profit, $\pi_1(p_1)$, together with the two period 2 profits—$\pi_2^H(2p_1)$ for the high-value segment and $\pi_2^L(p_1)$ for the low-value segment.[9] In our setup, the profit-maximizing price is $p_1 = \frac{3V}{10}$ and the threshold is $\tilde{v}_1(p_1) = \frac{6V}{10}$, such that the upper-40% of consumers, with values $v \in \left[\frac{6V}{10}, V\right]$, buy the good in period 1. Then, in period 2, the seller sets $p_2^H(p_1) = \frac{6V}{10}$ for the consumers who bought the product in period 1, such that the same consumers, with values $v \in \left[\frac{6V}{10}, V\right]$, buy also in period 2; and $p_2^L(p_1) = \frac{3V}{10}$ for the consumers who did not buy the product in period 1, such that consumers with values $v \in \left[\frac{3V}{10}, \frac{6V}{10}\right]$, buy in period 2. As compared to the II-IR case, we have fewer period 1 purchases and fewer period 2 purchases.

Comparison. What are the effects of BBP in the PI-PR case? Whereas the seller's profit was $\frac{1}{2}V$ without BBP, it is: $\pi_1(p_1) + \pi_2^H(2p_1) + \pi_2^L(p_1) = 0.45V$ with BBP. In terms of consumer surplus, as compared to a surplus of $0.25V$ without BBP, we have: $CS_1(p_1) + CS_2^H(2p_1) + CS_2^L(p_1) = 0.325V$ with BBP. When consumers are aware of the seller's use of BBP and respond strategically, BBP helps consumers and harms sellers. (This is why sellers would prefer to commit to refrain from using BBP, if they could.)

Drilling down further, we can distinguish between four groups of consumers, as shown in Table 3 below. The table also presents, for each group, the consumer surplus, the seller's profit and the total surplus (which combines the consumer surplus and the seller's profit), with and without BBP.

Consumers with	Consumer Surplus		Seller's Profit		Total	
	No BBP	BBP	No BBP	BBP	No BBP	BBP
$v \in \left[\frac{6V}{10}, V\right]$	$0.24V$	$0.28V$	$0.4V$	$0.36V$	$0.64V$	$0.64V$

[9] From the preceding paragraph, we know that: $\pi_1(p_1) = p_1 \cdot [1 - F(\tilde{v}_1(p_1))] = p_1 \cdot [1 - F(2p_1)]$, $\pi_2^H(p_2^H(p_1)) = p_2^H(p_1) \cdot [1 - F(p_2^H(p_1))] = 2p_1 \cdot [1 - F(2p_1)]$, and $\pi_2^L(p_2^L(p_1)) = p_2^L(p_1) \cdot [F(\tilde{v}_1(p_1)) - F(p_2^L(p_1))] = p_1 \cdot [F(2p_1) - F(p_1)]$. The seller sets a price that solves: $\max_{p_1}\{\pi_1(p_1) + \pi_2^H(p_2^H(p_1)) + \pi_2^L(p_2^L(p_1))\}$.

$v \in \left[\frac{V}{2}, \frac{6V}{10}\right]$	$0.01V$	$0.025V$	$0.1V$	$0.03V$	$0.11V$	$0.055V$
$v \in \left[\frac{3V}{10}, \frac{V}{2}\right]$	0	$0.02V$	0	$0.06V$	0	$0.08V$
$v \in \left[0, \frac{2V}{7}\right]$	0	0	0	0	0	0

Table 3: Disaggregated Effects of BBP in PI-PR Markets

We can now summarize the effect of BBP on each group: (1) Consumers with $v \in \left[0, \frac{3V}{10}\right]$ would be excluded from the market with and without BBP. (2) Consumers with $v \in \left[\frac{3V}{10}, \frac{V}{2}\right]$ would be excluded without BBP and served, albeit only in the second period, with BBP. (3) Consumers with $v \in \left[\frac{V}{2}, \frac{6V}{10}\right]$ would be served in both periods without BBP and only in the second period with BBP. Still, because of the lower price charged with BBP in the second period, they enjoy a higher consumer surplus; and the seller's profit from serving these consumers is lower. (4) Consumers with $v \in \left[\frac{6V}{10}, V\right]$ would be served, in both periods, with and without BBP. BBP allows the seller to charge a higher price in the second period, but pushes the price down in the first period. Overall, in group (4), BBP shifts surplus from the seller to consumers (the total surplus is not changed by the introduction of BBP). Looking across the four groups, BBP harms the seller and helps consumers; and, unlike in the II-IR case, all groups of consumers benefit.

C. Summary

In the PI-PR case, algorithmic behavior-based price discrimination is welfare enhancing, increasing both the consumer surplus and overall welfare. In the II-IR case, the welfare effects are more subtle. BBP reduces overall consumer surplus, but the harm is concentrated in the group of high-WTP consumers who are likely richer, whereas low-WTP consumers who are likely poorer benefit from BBP.

II. Algorithmic Quality Discrimination in II-IR Markets

Consider a market with two products, P1 and P2. The cost, to Seller, of manufacturing P1 is c_1 and the cost of manufacturing P2 is c_2. To focus on the effect of benefit and perceived benefit, we assume that $c_1 = c_2 \equiv c$.[10] Consumers enjoy a benefit b_1 from P1 and b_2 from P2; assume that $b_1 > b_2$.[11] We analyze two types of misperception:

(a) Overestimation: Biased consumers (mistakenly) think that the benefit from P2 is δb_2, where $\delta > 1$. For example, consider the market for new cars and assume, for simplicity, that there are two types of cars—one is larger with more leg-room and a bigger trunk (P1), whereas the other is smaller but comes with a higher-end entertainment system (P2). Consumers who overestimate the number of hours that they will spend listening to opera in the car will overestimate the benefit from P2. To focus on situations where the overestimation bias is potentially most troubling, we assume that $b_2 < b_1 < \delta b_2$, i.e., that the bias flips the relative desirability of the two products.

(b) Underestimation: Biased consumers (mistakenly) think that the benefit from P1 is δb_1, where $\delta < 1$. Consider, again, the market for new cars and assume, for simplicity, that there are two types of cars—one is a highly fuel-efficient hybrid vehicle (P1), whereas the other is much less fuel-efficient but comes with fancier seats and a higher-end entertainment system (P2). Since the benefit from P1 accrues over time, present biased consumers will underestimate this benefit. To focus on situations where the underestimation bias is potentially most troubling, we assume that $\delta b_1 < b_2 < b_1$, i.e., that the bias flips the relative desirability of the two products.

In both cases, we assume that a share α_1 of consumers are unbiased and recognize the true benefit (b_1 or b_2), whereas the remaining share $\alpha_2 (= 1 - \alpha_1)$ of consumers are biased and misperceive the benefit, as δb_2 in the overestimation case or as δb_1 in the underestimation case). Market power is such that Seller can set a price equal to a percentage $\gamma < 1$ of the consumers' benefit (or WTP).[12]

1. Overestimation

In a world with big data and sophisticated algorithms, Seller can distinguish between the biased and unbiased consumers, offering P1 to the unbiased consumers and P2 to the biased consumers.[13] In our example, the algorithm offers the larger vehicle to the unbiased consumers, at a price of $p_1 = \gamma b_1$. At the same time, the algorithm offers the smaller car with the high-end entertainment system to consumers who are identified as those who are likely to overestimate the benefit from the entertainment system. Moreover, the algorithm will set a high price for the smaller car with the high-end entertainment system, reflecting the biased consumers' inflated WTP: $p_2^B = \gamma \delta b_2$. Seller's overall profit, in an algorithmic world, is: $\pi^A = \alpha_1 (p_1 - c) + \alpha_2 (p_2^B - c) =$

[10] Consider relaxing the equal-cost assumption: $c_1 = c_2 \equiv c$.

[11] In a more general model, we would not assume a single benefit for each product, but rather two demand curves—one for each product.

[12] Alternatively, we could assume that the price leaves consumers with a share γ of the overall (perceived) surplus, e.g., $\gamma(b_2 - c)$. [Do as robustness check.]

[13] Seller will offer P1 to the unbiased consumers, because $p_1 - c > p_2^{UB} - c$, which is equivalent to $p_1 > p_2^{UB}$ or $\gamma b_1 > \gamma b_2$. Seller will offer P2 to the biased consumers, because $p_1 - c < p_2^B - c$, which is equivalent to $p_1 < p_2^B$ or $\gamma b_1 < \gamma \delta b_2$.

$\alpha_1(\gamma b_1 - c) + \alpha_2(\gamma\delta b_2 - c)$; and the overall consumer surplus is: $CS^A = \alpha_1(b_1 - p_1) + \alpha_2(b_2 - p_2^B) = \alpha_1(1-\gamma)b_1 + \alpha_2(1-\delta\gamma)b_2$.

To appreciate the potential algorithmic harm in such cases, we must compare the quality-discrimination outcome to the no-differentiation benchmark. What would car sellers do in a pre-algorithmic world, where they cannot distinguish the biased consumers from the unbiased consumers? Unable to discriminate, the sellers would offer the same car to all consumers.[14] But which car will they offer? Would they offer the larger car or the smaller? And what price will they set? The answer depends on market conditions—on the aggregate demand for each model, which depends on the number of biased vs. unbiased consumers.[15]

Which product will Seller offer—P1 or P2? If Seller offers P1, then misperception doesn't play a role (since only the benefit from P2 is overestimated). Seller sets a price of $p_1 = \gamma b_1$ and earns a profit of $\pi_1 = \gamma b_1 - c$. Note that all consumers buy P1. If Seller offers P2, then she must choose which consumers she wants to serve. If Seller wants to serve all consumers, she will set a price of $p_2^{UB} = \gamma b_2$ and earn a profit of $\pi_2^{UB} = \gamma b_2 - c$. Alternatively, Seller could forgo the business generated by the unbiased consumers and set a higher price, $p_2^B = \gamma\delta b_2$, at which only overestimators would make the purchase. Seller's profit will then be $\pi_2^B = \alpha_2(\gamma\delta b_2 - c)$, reflecting a higher per-unit profit but a smaller number of units sold. Therefore, in a pre-algorithmic world:

(i) Seller will offer P1, the larger car, to all consumers, if the profit that Seller can make from offering the larger car to all consumers exceeds the profit that she can make from offering the smaller car only to overestimators, i.e., if $\pi_1 > \pi_2^B$. In this case, consumer surplus will be $CS_1 = (1-\gamma)b_1$.

(ii) Seller will offer P2, the smaller car, at a price that will attract only biased consumers, if the profit that she can make from offering the smaller car only to overestimators exceeds the profit that Seller can make from offering the larger car to all consumers, i.e., if $\pi_2^B > \pi_1$. In this case, consumer surplus will be $CS_2^B = \alpha_2(1-\delta\gamma)b_2$.

Note that, since $\pi_1 > \pi_2^{UB}$, Seller will never offer P2 at a price that will attract all consumers. Intuitively, in order to sell the smaller car to all consumers, Seller would have to reduce the price to a level that even unbiased consumers would be willing to pay; Seller would not be able to price at a higher level that only biased consumers are willing to pay. But if such a low price is needed to capture the entire market with the smaller car, it is more profitable for Seller to capture the entire market with the larger car that can fetch a higher price.

To assess the welfare effects of algorithmic quality discrimination, we compare the pre- and post-algorithmic worlds. In case (i), quality discrimination harms consumers, since $CS^A < CS_1$. In a pre-algorithmic world, all consumers get the superior product (the larger car), P1, whereas

[14] We compare the option of offering only the larger vehicle or offering only the smaller vehicle. But there is another possibility: If sellers cannot discriminate, they might offer a third product design (i.e., not one of the two product designs described in the text). In this case, algorithmic discrimination might help some consumers while harming others.

[15] If we relax the equal-cost assumption ($c_1 = c_2 \equiv c$), then the answer will also depend on the relative manufacturing costs of the two models.

in the post-algorithmic world, the biased consumers get the inferior product (the smaller car), P2, and overpay for it. In contrast, in case (ii), quality discrimination helps consumers, since $CS^A > CS_2$. In a pre-algorithmic world, unbiased consumers are left out of the market, whereas in the post-algorithmic world, they get P1. (In both worlds, biased consumers get P2 and overpay for it.)

2. Underestimation

In a world with big data and sophisticated algorithms, Seller can distinguish between the biased and unbiased consumers, offering P1 to the unbiased consumers and P2 to the biased consumers.[16] In our example, the algorithm offers the hybrid vehicle to the unbiased consumers, at a price of $p_1^{UB} = \gamma b_1$. At the same time, the algorithm offers the low fuel-efficiency car to consumers who are identified as suffering from present bias, namely, to myopic consumers who fail to account for the significant long-term cost-saving that the hybrid vehicle promises; these consumers will be charged $p_2 = \gamma b_2$. Seller's overall profit, in an algorithmic world, is: $\pi^A = \alpha_1(p_1^{UB} - c) + \alpha_2(p_2 - c) = \alpha_1(\gamma b_1 - c) + \alpha_2(\gamma b_2 - c)$; and the overall consumer surplus is: $CS^A = \alpha_1(b_1 - p_1) + \alpha_2(b_2 - p_2) = \alpha_1(1-\gamma)b_1 + \alpha_2(1-\gamma)b_2$.

To appreciate the potential algorithmic harm in such cases, we must compare the quality-discrimination outcome to the no-differentiation benchmark. What would car sellers do in a pre-algorithmic world, where they cannot distinguish the present biased consumers from the unbiased consumers? Unable to discriminate, the sellers would offer the same car to all consumers. But which car will they offer? Would they offer the hybrid or the gas guzzler? And what price will they set? The answer depends on market conditions—on the aggregate demand for each model, which depends on the number of biased vs. unbiased consumers.[17]

Which product will Seller offer—P1 or P2? If Seller offers P2, then misperception doesn't play a role (since only the benefit from P1 is underestimated). Seller sets a price of $p_2 = \gamma b_2$ and earns a profit of $\pi_2 = \gamma b_2 - c$. Note that all consumers buy P2. If Seller offers P1, then she must choose which consumers she wants to serve. If Seller wants to serve all consumers, specifically if she wants to keep the underestimators, she will set a price of $p_1^B = \gamma\delta b_1$ and earn a profit of $\pi_1^B = \gamma\delta b_1 - c$. Alternatively, Seller could forgo the business generated by the biased consumers and set a higher price, $p_1^{UB} = \gamma b_1$, at which only unbiased consumers would make the purchase. Seller's profit will then be $\pi_1^{UB} = \alpha_1(\gamma b_1 - c)$, reflecting a higher per-unit profit but a smaller number of units sold. Therefore, in a pre-algorithmic world:

(i) Seller will offer P2, the gas guzzler, to all consumers, if the profit that Seller can make from offering the gas guzzler to all consumers exceeds the profit that she can make from offering the hybrid only to unbiased consumers, i.e., if $\pi_2 > \pi_1^{UB}$. In this case, consumer surplus will be $CS_2 = (1-\gamma)b_2$.

[16] Seller will offer P1 to the unbiased consumers, because $p_1^{UB} - c > p_2 - c$, which is equivalent to $p_1^{UB} > p_2$ or $\gamma b_1 > \gamma b_2$. Seller will offer P2 to the biased consumers, because $p_1^B - c < p_2 - c$, which is equivalent to $p_1^B < p_2$ or $\gamma\delta b_1 < \gamma b_2$.

[17] If we relax the equal-cost assumption ($c_1 = c_2 \equiv c$), then the answer will also depend on the relative manufacturing costs of the two models.

(ii) Seller will offer P1, the hybrid, at a price that will attract only unbiased consumer, if the profit that Seller can make from offering the hybrid to these unbiased consumers exceeds the profit that she can make from offering the gas guzzler to all consumers, i.e., if $\pi_1^{UB} > \pi_2$. In this case, consumer surplus will be $CS_1^{UB} = \alpha_1(1-\gamma)b_1$.

Note that, since $\pi_2 > \pi_1^B$, Seller will never offer P1 at a price that will attract all consumers. Intuitively, in order to sell the hybrid to all consumers, Seller would have to reduce the price to a level that even present-biased consumers would be willing to pay. But if such a low price is needed to capture the entire market with a hybrid, it is more profitable for Seller to capture the entire market with the gas guzzler that can fetch a higher price.

To assess the welfare effects of algorithmic quality discrimination, we compare the pre- and post-algorithmic worlds. In case (i), quality discrimination helps consumers, since $CS^A > CS_2$. In a pre-algorithmic world, all consumers get the inferior product, P2, whereas in the post-algorithmic world, the unbiased consumers get the better product, P1. Also in case (ii), quality discrimination helps consumers, since $CS^A > CS_1^{UB}$. In a pre-algorithmic world, biased consumers are left out of the market, whereas in the post-algorithmic world, they at least get P2 (which still provides a positive benefit).