Philosophy & Technology (2021) 34:967–992
https://doi.org/10.1007/s13347-021-00444-9

RESEARCH ARTICLE

Choosing how to discriminate: navigating ethical trade-offs in fair algorithmic design for the insurance sector

Michele Loi[1] · Markus Christen[2]

Received: 18 September 2019 / Accepted: 1 February 2021 / Published online: 13 March 2021

Abstract

Here, we provide an ethical analysis of discrimination in private insurance to guide the application of non-discriminatory algorithms for risk prediction in the insurance context. This addresses the need for ethical guidance of data-science experts, business managers, and regulators, proposing a framework of moral reasoning behind the choice of fairness goals for prediction-based decisions in the insurance domain. The reference to private insurance as a business practice is essential in our approach, because the consequences of discrimination and predictive inaccuracy in underwriting are different from those of using predictive algorithms in other sectors (e.g., medical diagnosis, sentencing). Here we focus on the trade-off in the extent to which one can pursue indirect non-discrimination versus predictive accuracy. The moral assessment of this trade-off is related to the context of application—to the consequences of inaccurate risk predictions in the insurance domain.

Keywords Fairness in machine learning · Fairness in insurance · Social justice · Egalitarianism · Prioritarianism · Utilitarianism

1 Introduction

Insurance has always been a data-driven business that relies on the statistical analysis of data about past cases and risk predictions regarding existing or prospective clients. Business models in the insurance sector are being innovated by the potential offered by new data, in particular that provided by global Tech and E-commerce companies as well

✉ Michele Loi
michele.loi@uzh.ch

[1] Institute of Biomedical Ethics and the History of Medicine, University of Zurich, Zürich, Switzerland

[2] Digital Society Initiative, University of Zurich, Zürich, Switzerland

as the Internet of Things sensors (including "trackers" and geo-localized devices). The sheer amount of new data available, combined with the sophisticated technologies available to analyze them, is contributing to an activity as old as insurance itself: predicting the individual risk of policyholders and charging them "actuarially fair" prices.

As data analysis (e.g., regression or machine learning algorithms) is applied to new types of data, the ethics of algorithms and the ethics of big data in insurance have begun to overlap. Algorithms can be used to assign a personalized premium. They can make or suggest decisions, for example, whether to reject a client or pay their claim. They justify such decisions based on predictions from the client's data, e.g., the prediction of a client's future cost, the probability of a fraud, or their willingness to switch to another provider.

In this paper, we provide an ethical analysis of discrimination in private insurance to guide the application of non-discriminatory algorithms for risk prediction in the insurance context. For the sake of simplicity, we focus on insurance decisions that depend on the magnitude of the risk insured against rather than the cost of claims processing and other factors influencing the customer's willingness to pay. The aim is to deliver non-comprehensive ethical criteria for data-science experts, business managers, and policy-makers (non-comprehensive in that they are based on a partial, preliminary mapping of the relevant ethical terrain). In doing so, we relate philosophical moral arguments to the anti-discrimination ("fair") algorithmic techniques proposed in the machine learning literature. The reference to private insurance as a business practice is essential in our approach, because we consider consequentialist arguments and the consequences of risk predictions for the insurance practice are different to those in other sectors (e.g., medical diagnosis, sentencing). Moreover, the computer science literature has demonstrated the existence of a trade-off in the extent to which one can pursue non-discrimination versus predictive accuracy, and the moral assessment of this trade-off is related to the context of application.

This essay does not aim at comprehensiveness, either in its review of the fair machine learning debate or in that of the ethical arguments potentially relevant to insurance. Questions of algorithmic transparency, interpretability, and accountability (Martin 2018) are also outside the scope of this paper. Rather, we build interdisciplinary connections between debates on discrimination and fairness in general across computer science and philosophy (Binns 2018; Custers et al. 2012; Gajane 2017) and those in the ethics of insurance. These debates have not yet been connected in the literature.

The paper has the following structure. Section 2 compares concepts of discrimination both internal and external to the insurance domain. Here, we focus on the concept of statistical discrimination, in both its direct and indirect forms, providing illustrations in the insurance domain. Section 3 maps the philosophical definitions from Section 2 onto the debate on fairness in machine learning. Section 4 presents ethical arguments against statistical discrimination, both direct and indirect. Section 5 considers the consequentialist arguments in favor of accurate statistical discrimination from the utilitarian, egalitarian, and prioritarian perspective, also considering traditional arguments from economics.[1] Finally, Section 6 illustrates two approaches with which to

[1] We define a moral view as consequentialist if and only if "the good is defined independently from the right, and then the right is defined as that which maximizes the good"(Rawls 1999, pp. 20–21). As a counterpart, we use the notion of "non-consequentialist," whereby deontological moral views are the most common family of moral theories that fall into this category.

combine all these arguments into a decision whether to use "fairer" (or better, less indirectly discriminatory) data-driven predictive tools.

2 The concept of discrimination

2.1 Discrimination in the insurance domain

Generally, the term "discrimination" can be used both in a purely descriptive sense (in the sense of making distinctions) or in a normative sense, which implies that differences in treating certain groups is morally wrong; the latter use of this term is more common in everyday speech. In the following, however, our use of "discrimination" does not imply that the unequal treatment is always or necessarily morally wrong (Lippert-Rasmussen 2014).

Our focus is on instances of unequal treatment that have a basis in sound statistical predictions, and we also use the word "discrimination" when this treatment is actuarially accurate.

More precisely, we restrict our investigation to *statistical* discrimination (definition 1), which involves treating a group worse (or better) than another group because of statistical evidence that the two groups differ in a dimension of interest (adapted from Lippert-Rasmussen 2007). In insurance, where differences in treatment are reflected by differences in premiums, such a situation would be described as "actuarially" fair, whereas "unfair discrimination" refers to the unequal treatment of individuals with the same risk level (Meyer 2004, p. 31). Unfair, in other words, means for insurers "equal risks are treated differently" (Meyer 2004, p. 31). According to this perspective, insurers do not unfairly discriminate when they charge a higher price to men than to women if actuarial data about car accidents suggests that the probability of claims of civil liability for damages is different between men and women. Charging men higher prices treats them worse than women because of the statistical difference between the two groups. In this case, the two groups are men and women, the dimension of interest is civil liability for damages, and the practice is charging higher prices to men because of the statistical evidence suggesting that they are more likely to be liable for such damages. Finally, our examples all involve the simplest possible statistical mathematical tasks and decision problems which may emerge in the insurance context. These are not fully realistic[2] but allow the reader to focus on the moral case for and against direct and indirect statistical discrimination.

2.2 Definitions of discrimination types

We now define the two relevant types of discrimination discussed in this paper. With reference to statistical discrimination, we define (definition 2) direct statistical discrimination as any instance of statistical discrimination in which information about membership of a group (which differs statistically in the dimension of interest from some

[2] For the sake of simplicity, we do not consider mathematical definitions of fairness for regression (Berk et al. 2017; Toon Calders et al. 2013; Komiyama et al. 2018); these definitions are conceptually and morally related to the simpler definitions discussed here.

other group) is used intentionally in the procedure that assigns better or worse treatment to an individual.

When we say that group information is used intentionally we do not mean by that, that the decision maker has a *preference* for or against the group, or even a *preference* for treating G1 *differently* for some other groups. We refer to "intentionality" in this sense as "weak," as opposed to the "strong" intentionality of someone who wants one group to receive worse treatment for its own sake. Weak intentionality indicates an intentional action of *enabling* membership to G1 to play, conceptualized *as* G1, a causal role[3] in the process leading to unequal decisions. This happens, for example, any time an insurer relies on a decision-making model involving sex as an independent variable in the context of insurance decisions. The decision to use a model involving sex to make decisions is an intentional act; therefore, the fact that information about sex plays a causal role in the decision-making process cannot be characterized as unintentional.

This type of intentionality is what makes discrimination direct, i.e., based on sex, not the fact that an insurer aims to treat individuals from different groups differently for its own sake (this would be, in our usage, strong intentionality). Rather, in weak intentionality, the insurer's aims may be different, such as pricing insurance products according to risk and/or maximizing profit.[4] Cases in which information about group membership is deliberately used to build a predictive model can be logically and morally distinguished from cases in which this information is used by the algorithm in a manner that does not reflect even the (weak) intentions of designers. This can happen when information about group membership is redundantly encoded in other data and the data scientist or pricing algorithm user is unaware of this.

Suppose, for instance, that a predictive model uses "short hair, does not purchase sunscreen, drinks no alcohol" to recommend lower premiums, which in a given population picks up Muslims 99% of the time. Informationally speaking, information about this combination of features can be considered information about religion in that population. We prefer to describe this as a case in which Muslims are indirectly favorably discriminated rather than one in which they are directly discriminated, as long as there is no awareness of the use of this information so the information is not used intentionally (in the weak sense). This distinction matters morally in terms of the responsibility of the insurer; for example, if insurers have a legal or moral duty not to treat people differently because of their race, insurers who (weakly) intentionally use

[3] When we say "causation," in this account, we assume an account of causation such that it is not necessary that there are natural laws that back up causation, as in nomological-deductive accounts. Our account of causation is more liberal in that it treats as a cause any relationship that one power could potentially exploit for purposes of manipulation and control of phenomena (Menzies and Beebee 2020; Woodward 2005). Neither does causation in this sense need to be deterministic.

[4] Other accounts of direct discrimination require preferences for or against certain salient social groups on the part of the decision-maker (Binns 2018). Not all do, however; for example, Lippert-Rasmussen proposes an even broader definition which treats group-related biases in assessing the evidence for a claim concerning an individual as a form of direct discrimination (Lippert-Rasmussen 2014, p.41). In US law, direct discrimination of groups indicated by anti-discrimination law roughly corresponds to the disparate treatment doctrine, according to which "classification itself is a legal harm, irrespective of the effect" (Barocas and Selbst 2016, p. 25). It seems to us that this is adequately understood as intentional classification, but not as classification serving a specific purpose. According to Barocas and Selbst, the idea of classification as an intrinsic legal harm includes "using protected class as an input to a system for which the entire purpose is to build a classificatory model" (Barocas and Selbst 2016, p. 25).

Muslim religion to make predictions may be in a different position morally or legally from insurers who use an almost perfect proxy with no awareness of this.[5]

The intuition we are attempting to capture through weak intentionality is the distinction between explicitly encoded and implicitly encoded information. However, the categories of explicit and implicit are not only vague but of unclear moral significance. An employer who uses ZIP codes as an (imperfect) proxy for race, i.e., to exclude applicants of a specific race, would engage in direct discrimination (even if the ultimate reason for direct discrimination was not an intrinsic aversion to a race but a desire to please their racist customers). It is also not possible to account for the difference between explicit and implicit information in terms of accuracy. All phenomena that can be fruitfully used to classify objects having a certain feature and that are not the feature itself can be considered statistical proxies of the feature that interests us. An application form asking customers to declare their race is as such no more intrinsically accurate than a proxy based on web browsing habits as, for example, people may deliberately lie about their race. Arguably, the law has yet to find a philosophically robust solution to distinguishing direct from indirect discrimination that addresses important questions with respect to the *aboutness* of information—what it means for information to be information *about* race. Philosophers therefore need to advance more precise and meaningful accounts when working with the distinction between direct and indirect discrimination.[6] Weak intentionality serves this purpose, at least as a working hypothesis.

Some of the groups discriminated against may be "socially salient." A group is perceived as such when "perceived membership of it is important to the structure of social interactions across a wide range of social contexts" (Lippert-Rasmussen 2007, p. 386). In our societies, feature types such as gender, race, ethnic group, or religion (especially when signalized by easily detectable features) are socially salient. That such groups are salient in our society is a contingent fact, not part of what "socially salient" means. With the exception of groups defined genetically, forms of discrimination considered problematic concern socially salient groups (Lippert-Rasmussen 2014). However, due to our interest in the big data context, in this paper, we also consider direct discrimination against groups that are not socially salient (e.g., purchasers of sport cars). Discrimination against non-socially salient groups may well become more widespread, affecting a greater number of individuals and life domains. In our use, every type of differential treatment for statistical reasons, including those not based on information about socially salient groups and not providing a proxy for socially salient groups, is an instance of direct statistical discrimination. As differential treatment for statistical reasons is likely to become more widespread and socially salient because of

[5] This is not meant to imply that the unintentional use of information that is discriminatory by proxy is morally neutral. In a context in which data scientists are supposed to be aware of the risks of redundant encoding, failing to check for correlations with protected groups may be regarded as negligence. For an analysis of the moral difference between the unintentional use of a perfect proxy and using explicit race information, see Tom Douglas, The Perfect Proxy Problem in Artificially Intelligent Crime Prediction, unpublished manuscript.

[6] Tom Douglas, in The Perfect Proxy Problem in Artificially Intelligent Crime Prediction, unpublished manuscript, suggests using the de dicto/de re distinction to distinguish implicit and explicit uses of information. By his own definition, the de dicto reading of discrimination based on race implies that the concept of race plays a causal role in the discrimination. However, a machine learning algorithm that has been intentionally designed to process ZIP information as a method to imperfectly distinguish race arguably also draws a racial distinction based on the human concept of race.

the widespread availability of data about people, direct statistical discrimination becomes more ethically important.

Let us now turn to indirect discrimination. Consider the following example: To assess the risk and premium paid for liability insurance, some UK car insurers rely on information concerning brand and type of the car, and whether the insured person has modified the car. Following an EU court judgment which established that using gender information to determine insurance premiums is against EU law (March 2011), the UK insurance industry cannot use information about gender. However, men still pay, on average, higher premiums than women (Collinson 2017). This is an example of the indirect discrimination of a socially salient group (and direct discrimination of non-socially salient groups).

In general, following Lippert-Rasmussen, the use of information about group G2 counts as indirect statistical discrimination (definition 3) against a socially salient group (G1-people) if and only if:

i) A non-socially salient group (G2-people) suffers from direct discrimination, and "it so happens that [G1]-people are more inclined to be [G2]-people than non-[G1]-people are"(Lippert-Rasmussen 2007, p. 389).
ii) G1-people on average, or most G1-people (Lippert-Rasmussen 2014),[7] receive the worse treatment as a result of this correlation.[8]

Referring to our example, G2-people are clients who have modified their cars, whereas G1-people are males. If male customers are more inclined to modify their cars than non-male customers, they will pay higher premiums on average, even if gender information is not collected directly.

In this paper, we consider indirect discrimination only against socially salient groups, following other philosophers (Lippert-Rasmussen 2014). An example of this is "redlining," which refers to denying a service (e.g., a loan) to residents of specific area. This counts as indirect discrimination against (racial or ethnic) minorities that reside predominantly in certain zones (Daniels 2004, p. 129).[9] Therefore, to return to the example, we do not consider indirect discrimination where an insurer uses data about whether clients subscribe to a car modification magazine—and are therefore likely to be car modifiers—in order to discriminate against car modifiers.

Note that the same case may be within the scope of our analysis qua case of direct discrimination but not qua case of indirect discrimination. For example, in the above

[7] Considering both group averages and most people in the group is required to classify as discrimination those unlikely cases in which most individuals in group are harmed by the criterion adopted, except a tiny minority who may be benefited, and the benefit for the tiny minority is so high as to mathematically neutralize all average effects on the group (Lippert-Rasmussen 2014).

[8] This definition of indirect discrimination is equivalent to disparate impact in US law. Disparate impact is not necessarily illegal, not even in those fields (e.g., employment) where there is law explicitly addressing it. Rather, disparate impact shifts the burden of the proof. When a plaintiff shows that particular facially neutral employment practice causes a disparate impact with respect to a protected class, "the defendant-employer may 'demonstrate that the challenged practice is job related for the position in question and consistent with business necessity.' Finally, if the defendant makes a successful showing to that effect, the plaintiff may show that the employer could have used an 'alternative employment practice' with less discriminatory results" (Barocas and Selbst 2016, p. 32).

[9] Assuming that area codes are not intended by insurers to be used as proxies of race information. This would make it direct discrimination according to the weak intentionality definition above.

example, it is within the scope of this paper to discuss moral reasons against the direct discrimination of car modification magazine subscribers, but not to discuss moral reasons against the indirect discrimination of car modifiers.

We do not assume that it is inherent to the concept of indirect discrimination that it only applies to socially salient groups. We merely assume that up to this moment in human history, most discrimination considered worth discussing has been against socially salient groups, with the noteworthy exception of genetic discrimination. This is far too ambiguous a social fact and, in our view, insufficient to treat such scope restriction as a norm of conceptual deployment. Suppose that a future algorithm widely used in the field of human resources treats differently individuals who drive red cars, eat rice, and have long hair. This combination of factors leads to sufficiently accurate predictions, but the trait is not socially salient; e.g., people with such features do not form social groups and are not distinguished in other social interactions. However, if the human resources algorithm is used pervasively, questions regarding the indirect discrimination against this group may become pertinent. For now though, considering all possible forms of indirect discrimination would turn the ethical analysis of statistical discrimination in insurance into an endless and impossible task, so we exclude that this may be a moral duty on pragmatic grounds.

In our paper, we do not explore the moral considerations concerning a third notion of discrimination, called disparate mistreatment. Nevertheless, as this topic has recently gained interest in the technical literature on discrimination in big data–based machine learning, we briefly introduce the term here. Disparate mistreatment can be defined as an inequality in the rate at which a socially salient group, G1-people, is liable to be misclassified by a statistical prediction tool and be treated disadvantageously, in comparison to another socially salient group (non-G1-people or a subgroup of them). It concerns, for example, the unequal likelihood of actually creditworthy clients from minority groups of being recognized and correctly classified as creditworthy. Discussing the moral difference between the two forms of disparate mistreatment is an extremely complex task that falls outside the scope of this paper (see, e.g., Zafar et al. 2017 for an in-depth discussion).

3 Algorithms against direct and indirect discrimination

Nowadays, computer algorithms are frequently used to develop (often complex) predictive models using (often large) quantities of data. In contemporary language, machine learning algorithms are said to "train" a "predictive model" with data. The input of machine learning algorithms are data (e.g., from past clients); the output are predictive models. These models can then be used to make predictions about new clients based on their data. Typically, algorithms are also used to recommend decisions based on such predictions. The decisions are based on the data of the new clients, e.g., their age, driving history, and type of auto driven, which, for the sake of simplicity in this analysis, we assume to have been labelled correctly, either by humans or automated systems.[10]

[10] This is not to imply that labelling issues are non-important for fairness and the ethical evaluation of such systems more broadly. In some fields, they may be one of the key drivers of unfair predictions (e.g., the confusion between the probability of arrest and the probability of crime may introduce significant biases in law enforcement or recidivism assessment) (for the importance of (mis)labelling, see Barocas and Selbst (2016)).

3.1 Avoid direct discrimination in machine learning

In the computer science literature, a predictive model satisfies "fairness through unawareness" if the algorithm "ignore[s] all protected attributes such as race, color, religion, gender, disability, or family status" (Hardt et al. 2016, p. 1). Given the above definition of direct discrimination (definition 2), this avoids direct discrimination relative to these groups. However, fairness as unawareness is no guarantee against indirect discrimination (definition 3). This is typically considered problematic in the computer science literature (Toon Calders and Verwer 2010; Kamishima et al. 2012; Pedreschi et al. 2008) as computer scientists have observed that correlations between socially salient groups (e.g., race) and non-socially salient groups (e.g., geographic zone, wealth) exist.

3.2 Algorithms against indirect discrimination

Computer scientists have developed algorithms which purportedly avoid indirect discrimination, which is equivalent to ensuring statistical independence between a protected (typically, socially salient) feature G1 (e.g., being a man) and the decision D (e.g., D could be the decision "reject the client"). The principle behind this is to ensure that any model using data which may be correlated with a socially salient group (e.g., wealth, purchases) will make decisions that do not correlate with membership to a socially salient group—e.g., women are less likely to receive credit. This can be achieved via several techniques, e.g., algorithms that reduce the decision weight of those traits that are more correlated with the socially salient group in the training and test datasets (T. Calders et al. 2009; Toon Calders and Verwer 2010; Feldman et al. 2015; Kamishima et al. 2012; Pedreschi et al. 2008).

3.3 Trade-offs

Using the algorithmic means described involves trade-offs. There is a trade-off between most definitions of fairness and accuracy. Intuitively, if fairness is defined as unawareness, then excluding information about group membership leads to predictions being less fine-grained when group membership is independently predictive of the dependent variable. If fairness is defined as statistical parity—equalizing the likelihood of a favorable decision between the different groups—then achieving fairness will typically lead to an accuracy loss even if no feature is excluded from consideration (Berk et al. 2018; Corbett-Davies et al. 2017; Dwork et al. 2012; Kleinberg et al. 2017). Authors that propose algorithms to avoid indirect discrimination typically suggest searching for a satisfactory compromise between fairness and accuracy, that is, reducing indirect discrimination to the point where an acceptable amount of data utility is lost.

Second, there are trade-offs between different fairness requirements. For example, if baseline differences in risk between man and women exists, requiring the same probability for man and women to be classified as low risk (statistical parity) worsens the gender imbalance in the false-positive or false-negative rates, which may also be viewed as unfair (Berk et al. 2018). Imposing statistical parity would also violate test-fairness in most cases;[11] thus, it is typically in tension with the fairness goal of avoiding

[11] When the base rates differ across groups, for imperfectly accurate predictions.

disparate mistreatment (see Section 2.2.). Discussing this pitfall falls outside the scope of this article.

A further trade-off occurs between not two but three fairness constraints, resulting in a trilemma. Typically, training a prediction tool to approximate statistical parity for a trait X (e.g., sex) causes a greater accuracy loss when the tool is not allowed to process information about X directly (Lipton et al. 2018). The trilemma can be stated as follows: You can achieve a fairness as statistical parity at the expense of accuracy, or fairness as unawareness at the expense of accuracy, or you may mitigate the trade-off between statistical parity and accuracy, but not without violating fairness as unawareness.

Finally, we note that an algorithm may be constructed that deliberately includes protected features (i.e., performs direct discrimination following our definition) to avoid indirect discrimination. This approach has been used, for example, in the context of affirmative action. This can be considered a case of justified direct discrimination for avoiding unjustified indirect discrimination, although it raises a well-known controversial discourse (Burns and Schapper 2008). A decision-making rule that allows group variables to affect a decision amounts to direct discrimination, given the definition of direct discrimination given here.

In summary, computer science has provided different mathematical definitions of non-discrimination in machine learning, but it lacks clear ethical guidance pertaining to their use. In what follows, we consider some philosophical arguments in ethics and political philosophy in favor or against the removal of direct and indirect discrimination.

4 Ethical arguments for why discrimination is morally objectionable

4.1 When is direct discrimination morally objectionable?

4.1.1 Fairness and choice

According to the choice principle, people ought not to be subjected to disadvantageous treatment because of something that does not reflect their own choices (Lippert-Rasmussen 2007, p. 398). So direct statistical discrimination is pro tanto morally wrong when an individual is imposed additional costs based on a feature G, where G is an unchosen trait. This may explain why discrimination based on gender, race, ethnicity, or genetics (Palmer 2007, p. 118) is widely perceived as morally problematic. Norman Daniels and others have argued that it is unfair to charge different rates based on traits outside the control of individuals, such as the genes with which people are born (Avraham et al. 2014; Daniels 2004, pp. 125–128).

The choice principle is not violated if people imposed additional costs as a consequence of their choice. However, in statistical discrimination, certain treatment that does not violate the choice principle can violate the "other people's choice principle" (Lippert-Rasmussen 2007). This is the principle that individuals should not be subjected to disadvantageous treatment because of something that reflects others' choices. Consider the example below:

> Example 2—shopping patterns and health risk: A statistical model is used to decide upon premiums based on their online shopping patterns. Those who

purchase a football outfit have higher risk scores and pay higher premiums for life insurance.[12]

Clearly, *purchasing an outfit* is a choice. However, suppose that the correlation between shorter life expectancy and football outfit purchases is driven by the unhealthy behavior of a group of football fans, who drink alcohol to excess and are more prone to be harmed or killed in riots. Now take Bob, who loves the aesthetic of football but is not a football fan and does not drink alcohol. Bob is subjected to disadvantageous treatment due to the lifestyle choices of football fans with whom he shares only a purchase pattern. The premium demanded of him reflects his choices (to purchase sport apparel) but also those of other people. It seems unfair to charge the football outfit purchaser a higher premium because of what other people choose to do (Lippert-Rasmussen 2007, p. 398).

Some cases of direct discrimination do not violate the other people's choice principle but violate the choice principle. Consider, for example, a life-insurer that charges higher premiums to clients who have a monogenic, high-penetrance disease. This practice does not violate the other people's choice principle because the likely effects of the gene on the client occur independently of others' actions. Some cases of direct discrimination violate neither the other people's choice principle nor the choice principle: Consider an insurer who treats smokers worse than non-smokers. The insurer subjects the smoker to higher premiums because of the smoker's choices. However, the smoker's risk will be higher because of the increased risk due to smoking, irrespective of what other smokers do.

Finally, it may be argued that individuals should not be subjected to disadvantageous treatment on the basis of their own choices if these choices reflect moral obligations and morally worthy choices (Thaysen and Albertsen 2017). One example could be the choice to live in a dangerous neighborhood to make a positive difference to the community, or forgoing a well-paid career to look after a loved one. Note, however, that as the moral worth of choices depends on the intention behind them, it is practically impossible, in most cases, for an insurer to determine the reason that affects the moral worth of individual choices.

4.1.2 The principle of avoiding decisions based on predictions reproducing injustice

What if the trait discriminated against is a socially salient group? In this case, there may be additional reasons against direct discrimination, beside those already discussed. According to Lippert-Rasmussen, statistical discrimination is morally objectionable, if the statistical facts conferring disadvantage on X_{G1}-people emerge as a result of morally objectionable social practices against X_{G1}-people, for which non-X_{G1}-people (or a subgroup thereof) are responsible (Lippert-Rasmussen 2007, p. 400). This is the principle that one ought to avoid supporting practices grounded in statistical generalizations that advantage oneself while disadvantaging others when the advantage only exists because of a morally unjustifiable treatment of the disadvantaged party by the advantaged party, which is what generates the social facts behind the statistics. We abbreviate this to the "principle of unjust statistical facts."

[12] Adapted from Kasper Lippert-Rasmussen (2007, p. 398)

The moral position behind this principle is that not all individuals are in the moral position to appeal to a certain kind of justification that grounds the unequal treatment of a group in an actual and relevant social fact concerning that group. The justification for the advantage—being grounded in statistical facts—would not pass what G.A. Cohen calls it the "interpersonal test" (Cohen 2008; Lippert-Rasmussen 2007, p. 401). The idea is that although the unjust treatment of the disadvantaged group may provide a justification that the disadvantaged group ought to accept, if the disadvantages were caused by other party, or sheer luck, then such a justification cannot be used by individuals who cause the social facts in question with their behavior when this behavior is both morally wrong and avoidable. To use this argument would be tantamount to denying one's responsibility for the injustice—to treat the social fact reflected in the statistics as a social outcome one could not help producing. This is illustrated by example 3 below.

> *Example 3—ethnic discrimination in dystopia: The country of Dystopia is inhabited by two different ethnic groups, "Asians" and "Caucasians." Asians have traditionally been richer, own most companies, and occupy most positions in the government. Until the last generation, Caucasians worked for them as cheap immigrant labor. Due to employment discrimination by Asian companies, Caucasians are poorer, more subject to unemployment, and are economically exploited in low-control working positions. This exposes them to poor health, with adverse effects on their cardiovascular risk and life expectancy (*Brunner and Marmot 2006*). In dystopia, life insurance companies use Caucasian ethnic membership as a proxy for shorter life expectancy and thus charge higher premiums, other things equal, to all clients whose ethnicity is Caucasian.*

According to the principle of unjust statistical facts, if it is morally objectionable for Caucasians to be exposed to racist treatment in the workplace, the statistical facts resulting from this cannot be cited by Asians in defense of the fairness of Caucasians paying higher premiums than theirs. Note that for a violation of the principle, the individuals causing the injustice and the individuals benefiting from it have to be the same individuals, not merely individuals belonging to the same groups. Admittedly, in a hypothetical world in which Muslims of generation x obtain economic advantages by discriminating Hindus, following which in generation $x + 1$ a mutual conversion occurs (all people born in Hindu families become Muslims and vice versa), it could still appear morally wrong for current Muslims (children of Hindus) to obtain lower prices. However, this intuition has nothing to do with Cohen's interpersonal test: If the new Muslims justify their advantageous treatment by appealing to statistical facts, they are not forced into the morally uncomfortable position of having to deny their own responsibility in unjustly causing those facts. On the other hand, we rely on Cohen's argument to explain the distinctive problematic nature of any justification for benefiting from statistical advantages associated with wrongly generated statistical facts because it is a relatively clear and simple argument. As the debate on reparation for historical injustices shows, there is no simple general argument which demonstrates that benefiting from historical injustice is a moral wrong that must be compensated as such (Perez 2011), and debating that large stream of literature falls outside the scope of this paper.

Historical injustice refers to the case in which all the original wrongdoers, and all the original victims, have passed away (Perez 2011). Therefore, we focus here on the relatively simpler case that is not one of historical injustice but of people benefiting indirectly from the injustice they have created.[13]

4.2 When is indirect discrimination morally objectionable?

Consider the example below.

> *Example 3 bis.—indirect discrimination of Caucasians in dystopia with big data: Everything is as in Ex. 3; that is, the shorter life span of Caucasians statistically justifies their higher premiums and is caused by widespread social injustices for which Asians are responsible. The main difference here is that insurance companies do not use Caucasian ethnic membership as a proxy for shorter life expectancy and thus do not charge higher premiums, other things equal, to all clients who are Caucasians. They predict life expectancy from shopping patterns; the predictive model infers a shorter life expectancy from purchases of sunscreen. In dystopia, most Asians do not like to sunbathe and stigmatize tanned skin. Thus, consumption of sunscreen correlates strongly with Caucasian ethnicity. Consequently, Caucasian clients pay 30% on average more for life insurance than Asian ones.* [14]

We will discuss whether the principles discussed in the previous section are violated by indirect discrimination by virtue of their rationale. Before we do so, it is useful to describe how the correlation between sunscreen purchases and higher prices arises via a causal diagram (Fig. 1).

This causal graph gives rise to a correlation between sunscreen purchases and life insurance prices at the end of two distinct causal chains. The correlation is due to the causal influence of race, which affects outcome through a social process that we deem unjust (direct workplace discrimination). Caucasian citizens are discriminated in the workplace by Asians because they are Caucasians, and if this had not occurred, they would not pay higher prices on average.

[13] The intuition of wrongness that Cohen's argument cannot explain can still be explained by showing that the example where the population benefiting from the statistical discrimination is "inverted" is also a violation of the other people's choice principle, so it may feel wrong for that reason as well.

[14] An example that violates only the unjust statistical facts principle, but not the choice principle, can be generated as follows: Suppose, for argument's sake, that a failure of institutions to maximize (through taxes and social services) the economic expectations of the worst off in society counts as unjust. Suppose that this social arrangement is unjust even in a hypothetical world in which all the worst-off citizens are worst off because of some imprudent choices they made, or their life-style preferences, which lead them to choose careers with lower incomes. Hence, the choice principle is not violated. We assume there that there is injustice even if the worst-off members of society are all worst off because of their choices and not because of their circumstances. (That such a society is unjust is implied by Rawls's view of justice as the inequality in question would violate Rawls's 1999 difference principle. This principle is agnostic with respect to individual choice and responsibility.) The worst-off group has a shorter life expectancy and characteristic patterns of consumptions, which are discovered through machine learning and lead to higher average premiums.

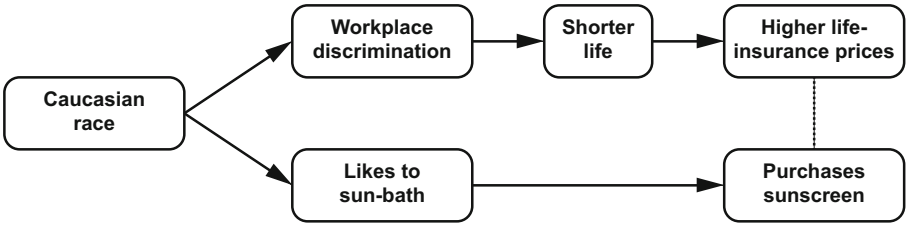

Fig. 1 Causal links responsible for correlations between sunscreen purchases and life insurance in Ex.3.bis

As race plays a causal role, the choice argument applies. The causal role of race is the relationship described in the graph: If Asians were not racists against Caucasians in the workplace, Caucasian lifespan would not be shorter than Asian lifespan (we assume this to be the case for argument's sake). Hence, we can say that insurers discriminate Caucasians because of their race even if insurers do not discriminate Caucasians based on their race. Caucasians are still discriminated because of their race (race plays a causal role),[15] although indirectly.[16]

There can also be cases of indirect discrimination in which a group is indirectly discriminated but group membership does not play a causal role, so individuals are not indirectly discriminated because of their race. Suppose that Caucasians simply happen, in a population, to be more inclined to purchase red cars. Then suppose that red car purchases are correlated with higher risk because there is some unknown psychological factor that causes clients to both purchase red cars and drive more dangerously. However, it is not the case that race is part of the best explanation for why Caucasians are more favorably disposed towards red cars than others: simply put, there is no explanation for this fact involving race. In other words, here the association between race and driving dangerously is not causally robust. No experiments have been conducted observing the variables under different circumstances, etc. No one has checked for a causal link, nor is there a plausible story for why this happens. The two variables are, let us suppose, merely correlated without a causal link, such as the correlation between the number of films Nicholas Cage appears in and number of people who drowned by falling into a pool, and many others (Vigen n.d.). We are ignorant of the causal relation, if any, and we are not in a position to check for causality. Therefore, no assumptions about causations are legitimate here. If the association were causal, i.e., race caused the purchase of red cars, the example would be similar to Ex.3.*ter* which we discuss below. In this case, one can say that Caucasians are more likely to receive worse treatment, so they are indirectly discriminated, but this does not happen because of race.

Let us now consider the other people's choice argument. This is also satisfied in the example 3 bis because Caucasians obtain higher prices because of other people's choices. Suppose that discrimination in dystopia is strong, but dystopia is not a caste society. Some Caucasians achieve the highest positions in society, especially as successful entrepreneurs.

[15] The causal role of race here is not conceived on race-realist terms. It be claimed that the effects that race causes are in fact constitutive of race as a social status, and give the social meaning of race. Thus, the analytical diagram concerning race needs a model of what constitutes race as a social phenomenon (Hu and Kohler-Hausmann 2020).

[16] In direct discrimination, Caucasians are discriminated based on race. As human-conceptualized race information directly influences the treatment (through a human action that intends to distinguish races, in a weak sense), Caucasians are also discriminated because of race.

Their children gain access to the best educational opportunities, healthcare, and jobs and have a high life expectancy, but their sunscreen purchasing patterns are indistinguishable from those of other Caucasians. For example, John is a high-status Caucasian. His individual expectation is to have a long life, but his insurer lacks the data points necessary to distinguish his case from low-status Caucasians. John is treated worse than Asians in the same social milieu because most Caucasians have social positions conducive to shorter life expectancy, which is something other Caucasians do: The other people choice principle is violated. (The same example can also be described as one in which John is violated because of his race [an unchosen trait], so the choice principle is violated, too.)

Let us now consider the unjust statistical facts principle. Indirect discrimination in dystopia is morally objectionable because it violates the principle of unjustly generated statistical facts, for similar reasons as direct discrimination. In this case, the statistical association between sunscreen purchases and lower life expectancy only occurs because of the unjust practices against Caucasians for which Asians are responsible. Therefore, Asians cannot use a statistical justification of their advantage against Caucasians. Example 3 bis thus violates all three principles in question. To clarify the difference between the different arguments, we now describe a case of indirect discrimination which violates only the choice and other people's choice principles:

> *Example 3 ter. Indirect discrimination violating the choice principle but not the unjust statistical facts principle. Suppose that lighter skin individuals are both more likely to develop skin cancer and more likely to be employed in manual work compared to darker skin ones. The shorter lifespan, let us suppose, is not in itself unjust. In this hypothetical world, workplace and other social opportunities are not influenced by objectionable social responses to skin color, such as racist attitudes. However, on average, having a lighter skin correlates with lower income. The predictive algorithm discriminates directly against people with lower income, which is statistically associated with lighter skin. This occurs because the machine learning algorithm learns to treat lower income as a proxy for lighter skin, which is what actually causes shorter lifespan (that matters to the insurer), through a purely biological mechanism which is fully understood. Lighter skin causes reduced life expectations via its molecular properties (causing both whiteness and reduced resistance to sun light). This makes lighter skin causally relevant, and the data correlated with it informationally relevant, for lifespan predictions and insurance costs. As a result of using income as a basis for insurance cost predictions, Caucasians are indirectly discriminated against.*

Just by looking at their effect, indirect discrimination in Ex3.ter is—let us assume—entirely indistinguishable from the indirect discrimination in Ex3.bis. In this case, the choice principle is violated because lighter skin plays a causal role in explaining why Caucasians pay higher prices: Caucasians would not pay higher prices if their skin were darker. The other people's choice argument is also violated because lighter skin people who limit their exposure to the sun and have a longer lifespan are also required to pay higher prices by virtue of their income data. The principle of unjust statistical facts is not violated because the advantage of people of lighter skin is not due to any morally objectionable social practice against the same people.

4.3 Overall assessment of reasons for using anti-discrimination techniques

In summary, direct and indirect discrimination against G1 can be deemed pro tanto morally wrong because of (at least) four non-consequentialist principles, when G1 is:

i) Any unchosen feature, such as sex (4.1.1) (choice principle)
ii) Any chosen or unchosen feature G1 that make the client's risk assessment dependent on the choices of others (4.1.1) (other people's choice principle)
iii) Any chosen feature G1 reflecting morally worthy choices, which are beneficial for society (worthy choice principle)
iv) Any chosen or unchosen feature G1 indicating a socially salient group where the predictive value of, and indirectly discriminatory effect on, such group is caused by morally objectionable practice against G1 (4.1.2), of which members of G2 are responsible, where members of G2 benefit from the statistical facts produced by such practice, and where the individuals benefiting from and responsible for the injustice are the same (unjust statistical facts principle).

5 Moral reasons in favor of accurate predictive models

In the section above, we examined the moral reasons to avoid direct and indirect discrimination. However, eliminating direct and indirect discrimination typically delivers less accurate predictive models (see Section 3.3). This section explores the moral implications of this trade-off by focusing on the positive reasons to value more accurate predictive models. Here, we assume that it is generally good for society that people are insured against risk. We consider the standpoint of insurance companies and regulators who want to maximize the social benefit, in a manner compatible with the long-term economic sustainability of the insurance businesses. We also assume that there is no willingness by governments to replace private insurers, turning a private insurance market into a no-longer-voluntary scheme of social insurance. The reason for this assumption is that moral and political arguments for or against social insurance fall outside the scope of this paper. Eventually, we will combine these arguments in favor of statistical discrimination with the previous arguments against statistical discrimination into a unified moral framework which aims to provide guidance to companies and regulators.

5.1 Adverse selection

Adverse selection is most often considered a justification for insurance companies to discriminate based on the most accurate feasible assessment of a client's risk. To illustrate the problem, consider a company selling a life insurance product and charging the same premium to all its clients, ignoring differences in the individual risks associated with gender and medical history. The product will attract more unhealthy men and less healthy women as the former have more to gain from purchasing such a product than the latter. The resulting pool will be costlier to insure than a random sample from the population. The high premiums necessary to insure a more high-risk pool further discourage the participation of healthy

women, driving the costs of insuring the pool (and the premiums) further up and further discouraging low-risk clients from joining the pool. This results, after a few iterations of this process, in a pool that includes none but the highest risk clients, which is very difficult or impossible to insure.

Note that these economic arguments do not refer to solidarity insurance (also called social insurance). In solidarity insurance, losses are paid to all, in equal amounts or on the basis of need; and each person pays the same, on the basis of need, or according to some other standard, but not in proportion to risk. Compulsory social insurance will be stable because low-risk individuals cannot leave the pool; large groups of citizens are legally obliged to buy insurance at a price defined by the regulator (Wilkie 1997, p. 1042). This averts the threat of adverse selection. Some may see social insurance as the best solution to the problem of fairness and discrimination in insurance (O'Neill 2006). However, it seems unreasonable to make every insurance product compulsory, at prices defined by regulators, especially if insurance does not meet a basic need (in contrast to products for which the socialized solution seems attractive, such as, arguably, healthcare). Here, we deal instead with so-called mutuality insurance which, by definition, is sustained in the market by the free choices of the insured, motivated by self-interest (Wilkie 1997, p. 1042). It may be objected that this creates a slightly artificial divide between private and social insurance markets, whereby the former operates according to a free market logic and the latter serves policy goals which make it less profitable or run at a loss. In reality, all markets are social, in the sense they depend on institutional and legal structures, including rights and freedoms, rule of law, and social protections. However, price regulation alone may not solve the adverse selection problem for non-compulsory insurance because the low-risk individuals may just decide not to purchase insurance at the price set by regulators. It is morally difficult to justify the coercion of low-risk individuals to satisfy the preferences of high-risk clients for insurance that does not meet a basic need (Scanlon 1975).

Adverse selection, however, is not guaranteed to occur whenever insurers do not use all the information available to them in principle. If low-risk policyholders do not know their risk level, are not perfectly rational, or lack alternatives, an insurance scheme may be sustainable in the long term, even if it requires low-risk types to implicitly subsidize high-risk ones (Heath 2007, p. 156). Empirical research shows that the threat of adverse selection is contingent on specific features of the particular insurance market in question (Avraham et al. 2014, p. 205). One implication of this is that insurance companies are not obliged to use risk classes corresponding to the smallest possible groups of homogeneous risk. In practice, they do not: If the cost of collecting risk information is excessive, insurers will not use it (Palmer 2007, p. 120).[17] Insurance schemes may survive even in the presence of implicit cross-subsidization between high- and low-risk groups, especially if, due to a legal prohibition against the use of certain risk information, no competitor can identify and "steal" low risk clients (Joseph Heath 2007; Palmer 2007). However, in some cases, adverse selection may occur in the absence of competitors because low-risk individuals do not find insurance at average community price sufficiently attractive and thus prefer to remain uninsured (this, however, can be avoided by making insurance compulsory). To summarize, the threat of adverse selection has variable plausibility, depending on features of the context (especially regulation) that we have identified.

When the threat of adverse selection is plausible, there are utilitarian reasons for society as a whole to avoid it. For any non-compulsory insurance *I*, a society where some people obtain *I* at higher prices and other people obtain *I* at lower prices is Pareto superior to one in which no one has *I*. Compared to a society without *I*, all people who are willing to purchase *I* at its market price are better off and no one is worse off.[18] Hence, there are utilitarian reasons to prevent insurance schemes from being eroded by adverse selection and become unstable. This counts against using more inaccurate non-discriminatory predictive models that expose the insurer to adverse selection effects. Note, however, that strict egalitarian consequentialism provides prima facie reasons against a society in which some people can afford insurance products and some cannot. We consider egalitarianism in 5.3, where we show that the assessment of insurance inequality from an egalitarian perspective is more complicated than the last claim suggests.

5.2 Incentives

Another argument in favor of risk-based discrimination in the insurance context concerns the value of calculating risk appropriately as part of a scheme of economic incentives to avoid or reduce such risk. This can, again, be reconstructed as an argument concerning the higher efficiency of arrangements where individual premiums are allowed to reflect individual risk. Economists also refer to "moral hazard" in relation to a "negative" version of the incentive argument, pointing out that insured people tend to increase their risk unless there are counter-incentives (Avraham et al. 2014, p. 206). However, the incentive argument presupposes that:

1) The policyholder has control over one or more variables (e.g., behavioral variables) of the risk insured against.
2) The policyholder can adjust their behavior on the basis of information (at a minimum, price information, known to be proportionate to the risk variables in 1) provided by the insurer.
3) Economic incentives effectively motivate the client to reduce their individual risk level for the risk insured against.

Again, these reasons count against using algorithmic techniques against indirect discrimination (see 3.2.) that have a high accuracy cost. For example, in the case of car liability insurance, inaccurate models may send confusing signals to the drivers which compromise the effectiveness of monetary incentives.

5.3 Outcome equality (and priority)

Utilitarianism assigns the same value to every event producing equal utility, assigning no intrinsic moral value to the way in which utility is distributed. In other words, there are always impersonal moral reasons to promote a larger aggregate of utility as an

[18] That is, the first, unequal outcome contains some people who are better off, and no one who is worse off, compared to the second, equal outcome. This, of course, assumes that the uninsured are not worse off due to envy or the indirect impact of lacking insurance, in zero-sum competitive contexts (Brighouse and Swift 2006).

alternative to a smaller amount, even if the only method to achieve this is to distribute utility more unequally among individuals. By contrast, egalitarianism attaches intrinsic moral value to a more equal distribution of utility[19], and prioritarianism attributes higher intrinsic moral value to utility, the lower the utility level of the person whose utility it is (Parfit 2003). Egalitarianism and prioritarianism provide different moral lenses to examine the questions of incentives and adverse selection. They capture two different ways of being "egalitarian" in the political sense: caring about equality and caring about the worst off in society.

There are utilitarian reasons to favor risk-based discrimination if it is part of a system of incentives leading clients to reduce their risk exposure. Incentives, however, may contribute to worsening inequality. Egalitarianism counts against efficient incentives that increase inequality in society, as in the example below:

> *Example 4—incentives and inequality in dystopia*
> *In dystopia, fitness trackers are introduced to measure individual physical exercise, and prices are lowered for the clients who adopt the healthiest lifestyles. It turns out that high-wealth people exercise more and low-wealth people less. Before the introduction of incentives, low-wealth clients paid on average 10% more than low-wealth ones for health insurance. After the incentives, they pay 30% more.*

The inequality described in Ex. 4 is not necessarily problematic from a prioritarian perspective. For example, prioritarians would favor incentives in the scenario described by Ex. 4 *bis*.

> *Example 4 bis—incentives and inequality in dystopia with significant savings: Everything is identical to Ex. 4. Because of overall savings produced by incentives, both low-wealth and high-wealth clients pay on average less in absolute terms compared to what they would pay without incentives.*

Moreover, a prioritarian theory attaches less value to the utility of better-off individuals, which is not the same as no value. Sufficiently large benefits to a large group of better-off individuals may outweigh, from a prioritarian perspective, the small benefits for a small group of individuals in the worst-off group. Thus, prioritarianism (but not egalitarianism) would be against the elimination of incentives in Ex. 4 *ter* below:

> *Example 4 ter—anti-discrimination law in Utopia*: *Utopia implements risk-insensitive insurance pricing, providing no incentive to avoid risk. As a result of eliminating existing incentives, prices grow 100% for all, except the least fit client, who obtains price reductions by 10%, favoring predominantly clients from the economically worst-off group. A 10% savings on insurance costs does not significantly improve the lives of citizens in the worst-off group.*

[19] This is ordinarily called "telic" egalitarianism (Parfit 2003).

Finally, note that most of the influential self-described egalitarians, for example, Jerry Cohen and Lerry Temkin, are value pluralist. This means that while defining justice (or fairness) as equality,[20] they also consider justice (or fairness) as one value among others (a very important one, but not always or necessarily an overriding one) to be considered in the design or evaluation of policy. Thus, even egalitarians could agree with prioritarians and utilitarians that the anti-discrimination policy proposed in Utopia is ultimately undesirable.

Prioritarianism, like utilitarianism, supports incentives if they benefit persons in the worst-off group in absolute terms (e.g., because the total cost of insuring the pool decreases, which allows everyone's premiums to be lowered to some extent), even if better-off clients benefit from the premium reduction, proportionally, more than the worst off. It even supports incentives where the worst-off group is made only slightly worse off as a result, whereas the majority of individuals are significantly benefited by them.

The problem of adverse selection (5.1) can also be assessed from an egalitarian and prioritarian perspective. Prioritarians have, in general, good reasons to avoid adverse selection as all individuals, both those who pay more and those that they pay less for their premiums, will be worst off in absolute terms without the kind of insurance they can and choose to afford (in an uncoerced setting).

Even strict egalitarians have reasons to favor a stable insurance scheme, as such, compared to a world without insurance, even when the premium paid to the insurer is the highest for the already worst-off group. Egalitarian reasons become clearer if one considers the distinction between ex ante inequality, that is, inequality before insurance claims are paid, and *ex post* inequality, that is, inequality after insurance claims are paid (Durnin et al. 2012). Imagine a society divided between poor villagers who live in seismic countryside zones and richer citizens that live in non-seismic urban zones. Poor villagers pay more than citizens to insure their homes. Let us assume for the sake of argument that the unequal contribution to insurance (higher for the poorest members of society) contributes to wealth inequality in society compared to a world in which insurance premiums are equal between seismic and non-seismic zones. Now suppose that an insurance scheme where premiums for seismic and non-seismic zones are equal is not sustainable in the long term due to adverse selection issues. Consider the possibility of no one having access to insurance that, in this scenario, is the only egalitarian alternative to clients paying different prices based on their level of risk (which entails that the poorest members of society pay higher premiums). From the ex post perspective, no insurance leads to more wealth inequality compared with the scenario of private insurance at market (risk-sensitive) prices. If no one is insured, an earthquake that destroys all the villagers' homes, leaving those in the cities intact, will further exacerbate the existing inequality between villagers and citizens. If the poor have housing insurance, even if they pay more for it, the claims paid by insurance mitigate the ex post inequality after the insured event has taken place. From the ex post

[20] Temkin and Cohen both favor a version of luck egalitarianism; namely, they identify justice (or fairness) with equality relative to outcomes that result from comparable choices. However, this is not really the point here, for philosophers like Temkin and Cohen tend to be very generous in assigning inequalities to circumstances rather than choice. In relation to Example 3, they could claim that the comparatively lower disposition to adopt healthier lifestyles of disadvantaged sectors of the population cannot often be considered a matter of personal responsibility and explain the pattern by pointing out the different circumstances experienced by low-wealth and high-wealth persons.

perspective, strict egalitarian consequentialists have egalitarian reasons to prefer that everyone be insured at equal premiums, but if this is not economically sustainable, they have the most reasons to prefer that everyone be insured at market rates compared to no private insurance being available.

This is, again, relevant when considering the utility costs of inaccurate predictive models. If predictive models that do not indirectly discriminate against previously disadvantaged groups are too inaccurate, the companies using them may not be economically sustainable. In this case, there will typically be utilitarian and prioritarian reasons against using such predictive models. Moreover, there will also be egalitarian (consequentialist) reasons against using them, from the ex post perspective, provided that the poorest high-risk clients can afford paying the higher premiums charged to them and—although losing some wealth because of their high premiums—are not excluded from insurance altogether.

6 Morally acceptable inaccuracy

Let us now review and combine together in single ethical framework moral reasons for and against mitigating discrimination while reducing accuracy. We emphasize that this is a modest proposal which sketches how the implications of the different moral principles considered in this paper could be combined into a decision framework. We do not claim that the methodology is immediately applicable, for that would require empirical testing and a careful analysis of the plausibility of its epistemic and pragmatic requirements. With this contribution, we hope to initiate a debate about the type of moral reasoning that would be appropriate for real-world agents to use make such decisions, in light of the existing degree of moral disagreement and uncertainty regarding the validity of philosophical moral theories.

Here, we sketch two methodologies of moral analysis involving heterogeneous moral premises. The heterogeneous premises form two groups: (1) the pro tanto reasons to avoid discrimination against G1, provided by the choice principle, the other people's choice principle, the worthy choice principle, and the unjust statistical facts principle, and (2) consequentialist reasons that evaluate the effects of accurate or less accurate risk predictions. The general framework we propose is the following: When indirect discrimination against a trait violates any of the above non-consequentialist principles, insurers have a pro tanto reason to use machine learning to produce a less discriminatory algorithm. Removing indirect discrimination entirely may not, however, be the all things considered morally required action because this has an accuracy cost which may have consequences, such as interfering with incentives and causing adverse selection which, in turn, influence the distribution of benefits and harms in society. These benefits and harms, however, can be evaluated from a distinct normative perspective. Typically, the utilitarian, prioritarian, and egalitarian moral assessment of such consequences will not be the same.

A simple approach would be to determine which of the non-consequentialist and consequentialist principles, if any, are correct. However, ultimate answers to these philosophical questions are not realistically forthcoming; in fact, they may represent rationally irresolvable disagreement. We present here two methodologies for dealing with our inability, as philosophical ethicists, to identify whether any principle or theory

invoked here is the, or one of the, valid options. The first method, expected choiceworthiness (MacAskill and Ord 2020), frames moral uncertainty (uncertainty about the normative validity of a principle or theory) in a way analogous to empirical uncertainty. The second approach treats moral disagreement as a practical political problem of providing reasons to support the same choices or policies for people who have different ultimate views about morality (Rawls 1996). The two approaches are not mutually exclusive, but, for simplicity's sake, we apply the second approach while assuming that different stakeholders maintain their views with certainty and each stakeholder only maintains one view.

The first approach is maximizing expected choiceworthiness (MacAskill and Ord 2020). Here, the decision-maker (e.g., the insurer) must first assign a credence value to the truth of each moral principle bearing on the question. These numbers are purely subjective and cannot be precise, nor do they have to be (MacAskill and Ord 2020). For instance, an insurer could assign the following credence score to the moral principles discussed in this paper: choice principle $p = 0.5$, other people's choice principle $p = 0.3$, unjustly generated statistical facts $p = 0.7$, utilitarianism $p = 0.3$, prioritarianism $p = 0.5$, and egalitarianism $p = 0.2$. (We shall ignore the worthy choice principle in this case: It is irrelevant because sex is normally unchosen.) Subsequently, the moral decision-maker should attribute a choiceworthiness value to each feasible practical choice, which expresses the desirability of that choice (i.e., the reasons in favor of it) from the standpoint of each principle in question. For the insurers in our example, the different choices are the different algorithms that will be used to assign a premium for clients. To illustrate, let us consider three of them: A1, which is the most accurate but also the one with the highest degree of indirect discrimination against G1; A2, a very inaccurate predictive model avoiding indirect discrimination; and A3, which has intermediate accuracy and indirect discrimination compared to A1 and A2. We assume that G1 is sex and that any degree of indirect discrimination against it implies a violation of the choice and unjust statistical facts principles, but not the other people's choice principle. The choiceworthiness scores in the table indicate degrees of choiceworthiness: i.e., strength of reasons in favor or against using each algorithm to assess and price risk. A small difference in the numbers indicates that one option is slightly more choice-worthy than another; a big difference indicates that there are strong reasons to choose in favor of an option and against the other. A negative value counts as a reason against and a positive value as a reason in favor. We also assume that one can make approximate quantitative comparisons of choice-worthiness across these different principles and theories (MacAskill and Ord 2020). The precision of the numbers should not mislead people into assuming that the method is more precise than it actually is. The probabilities are merely subjective probabilities, i.e., degrees of credence, and they can be produced by whatever method is deemed acceptable to produce subjective probability values, where moral uncertainty is not at stake. The choiceworthiness values can be generated by asking experts to indicate, within a given reference interval, how strongly morally desirable/undesirable a certain action or outcome would be according to the theory under consideration, assumed to be correct, in comparison to how it would be, according to a different theory, assumed to be correct.

For ease of exposition, we consider here only the relevant moral reasons, ignoring the prudential reasons of the insurer (e.g., return to shareholders, strategic advantages) that also have weight in determining all-things-considered expected choiceworthiness in combination with moral reasons (Table 1).

Table 1 Moral choiceworthiness scores of three insurance algorithms from the perspective of six different moral principles. The interpretation of the numbers in the table is provided in the paragraphs below

Algorithm	Choice	Other people's choice	Unjust statistical facts	Utilitarianism	Prioritarianism	Egalitarianism
A1	-100	0	-100	100	30	-30
A2	0	0	0	0	0	0
A3	-50	0	-50	80	90	-10

The choiceworthiness scores in the table can be interpreted as follows:

- A1 and A3 are both objectionable because they both involve a violation of the choice and unjust statistical facts principles, but not the other people's choice principle. But A1 is more objectionable than A3 because the inequality between the two sexes is higher.
- A1 produces very accurate predictions and it insures the most clients, so it produces the highest aggregate utility. Prioritarianism weakly supports this solution because these clients are predominantly already well off. Egalitarianism ascribes a negative value to A1, because it makes the world more unequal as the clients of the sex discriminated against are also predominantly poorer than the clients favorably discriminated.
- A2 avoids indirect discrimination, but it is not feasible in free market conditions, so its consequences are those of a world without such insurance. The strength of consequentialist reasons in its favor amounts to 0 from the perspective of each theory, because implementing A2 does not make the world a better place in terms of aggregate utility or in terms of utility for the worst-off individuals, and it does not make society more equal but neither does it make it worse.
- A3 is a predictive algorithm generating personalized prices that indirectly discriminates against G1 but less so than A1. The strength of utilitarian reasons in favor of A3 is weaker than that in favor of A1, because A1 insures and benefits significantly fewer clients than A3. From the prioritarian perspective, reasons in favor of A3 are stronger, because there are more clients from the worst-off group. Finally, there are egalitarian reasons against A3, because members of the sex discriminated against are on average poorer, which exacerbates existing inequalities, even if less so than A1.

The expected moral choiceworthiness of each option is the weighted sum of the choiceworthiness of each option from the perspective of each moral principle, and where the weight corresponds to the credence of the decision-maker in each principle. Hence, we have:

$$A1 = 0.5(-100) + 0.7(-100) + 0.3(100) + 0.5(30) + 0.2(-30) = -81$$
$$A2 = 0.5(-0) + 0.7(0) + 0.3(0) + 0.5(0) + 0.2(0) = 0$$
$$A3 = 0.5(-50) + 0.7(-50) + 0.3(80) + 0.5(90) + 0.2(-10) = 7$$

Based on this calculation, the most choiceworthy algorithm for this decision-maker in this hypothetical case is A3.

In providing an example of the second approach, we assume that each stakeholder is fully certain about the validity of one principle, for the sake of simplicity. The question for the decision-maker is whether it can abandon the most accurate predictive model and actuarially fair prices for an algorithm that is less indirectly discriminatory against a given sex, in a manner that appears legitimate to all its clients. We follow Rawls's (1996) strategy of searching for a justification that could achieve an overlapping consensus between individuals who differ in their ultimate moral views. For the sake of illustration, we assume that all clients believe that the indirect discrimination of one sex is pro tanto morally objectionable and that these clients are, in variable proportions, utilitarian, prioritarian, or egalitarian.

An overlapping consensus rule for adopting the algorithm *x* is:

Adopt Ax in place of the existing status quo algorithm A1, if and only if:

a) Ax generates higher aggregate utility than A1.
b) Ax improves the conditions of the worst off in society, relative to A1.
c) Ax generates less social inequality than A1.
d) Ax is sustainable (not eroded by adverse selection) in free market conditions.

Utilitarians, prioritarians. and egalitarians all have reasons to prefer Ax to A1 when this rule is satisfied, even if Ax is not the optimal algorithm from the perspective of any single principle.[21] This rule and the decision that follows from it are supported by an overlapping consensus of different moral doctrines.

This is achievable because each criterion promotes value as described by each principle without maximizing this value. Ax could achieve somewhat lower prices for the group paying the highest prices in relative terms, compared to the status quo, and somewhat higher prices for the group paying the lowest prices in relative terms; in this manner, it could produce more utility than A1 by covering more clients from the group paying the highest prices (even if it loses a greater proportion of more profitable clients from the advantaged group), and it would benefit more clients in the worst-off group, mitigating average inequalities. The resulting insurance pool could avoid adverse selection, satisfying (d), if, for example, a sufficient number of low-risk insurance clients are willing to cross-subsidize more high-risk clients for ethical, solidaristic reasons and accept paying slightly higher prices than competitors. This algorithm may not maximize the insurer's profit, but it could be deemed to be ethically superior from different ethical perspectives.

In some cases, however, no algorithm will be found that satisfies the overlapping consensus rule. When this occurs, the overlapping consensus approach provides no legitimation for the insurer to avoid using the accurate predictive model and actuarially appropriate prices.

7 Conclusion

In conclusion, predictive algorithms based on big data, which are used to assess risk (and adjust premiums accordingly), can lead to direct discrimination, indirect

[21] In keeping with Rawls' (1996) approach, these utilitarians, prioritarians, and egalitarians are assumed to be reasonable in the sense that they know that they must agree on a common rule binding for all and that others cannot be (realistically) persuaded or (due to moral constraints) forced to change their moral views.

discrimination, or the disparate mistreatment of certain groups. The present paper identifies some philosophical moral arguments concerning the moral wrongness of statistical discrimination which support avoiding direct and indirect discrimination against specific groups. Nevertheless, there is always a cost to avoiding indirect discrimination (by omitting to use some type of information or using more sophisticated machine learning methods) in terms of the reduced accuracy of the predictions. This causes further economic effects, e.g., undermining incentives for risk reduction, which are evaluated differently by utilitarians, prioritarians, and egalitarians. Finally, our paper proposes two approaches for bringing together the moral assessments produced from such diverse standpoints into some kind of decision regarding the most preferable algorithm.

Finally, we remind the reader that there may be other moral reasons for or against certain methods of using big data for improving predictive models (e.g., because certain types of data involve privacy risks) that may need to be considered in a more holistic evaluation in certain settings. Furthermore, other applications of big data analytics in insurance, such as personalizing claims processing or pricing based on willingness to pay, may raise additional ethical issues not covered by this analysis. However, we hope that our approach can be developed further with the help of stakeholders and inspire similar approaches in other practices where machine learning is used for predicting risk.

Acknowledgements Far too many people have had opportunities to provide feedback on different versions of this article, which has been in gestation since 2016, for us to list them all here. The authors wish to thank in particular Dr. Tim Räz for his remarks on causality and Dr. Benno Keller for his feedback on insurance business models, as well as two anonymous referees of the Philosophy of Technology and three anonymous referees of the Journal of Business Ethics.

Funding Open Access funding provided by Universität Zürich. This study was funded by the Swiss National Science Foundation, grant number 407540_167218 / 1.

Declarations

Conflict of interest The authors declare no competing interests.

References

Avraham, R., Logue, K., & Schwarcz, D. (2014). Understanding insurance antidiscrimination laws. *Southern California Law Review, 87*(195) https://scholarship.law.umn.edu/faculty_articles/576.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(671), 671–732.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2017). A convex framework for fair regression. *ArXiv:1706.02409 [Cs, Stat]* http://arxiv.org/abs/1706.02409.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: the state of the art. *Sociological Methods & Research, 0049124118782533*. https://doi.org/10.1177/0049124118782533.

Binns, R. D. P. (2018). Fairness in machine learning: lessons from political philosophy. *Journal of Machine Learning Research*. https://ora.ox.ac.uk/objects/uuid:2ff2785b-b0d4-447a-8326-a1fcc4c80840.

Brighouse, H., & Swift, A. (2006). Equality, priority, and positional goods. *Ethics, 116*(3), 471–497.

Brunner, E., & Marmot, M. G. (2006). Social organization, stress, and health. In R. G. Wilkinson & M. G. Marmot (Eds.), *Social determinants of health (2nd ed., pp. 6–30)*. Oxford University Press.

Burns, P., & Schapper, J. (2008). The ethical case for affirmative action. *Journal of Business Ethics, 83*(3).

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery, 21*(2), 277–292. https://doi.org/10.1007/s10618-010-0190-x.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. *IEEE international conference on data mining workshops, 2009*, 13–18. https://doi.org/10.1109/ICDMW.2009.83.

Calders, T., Karim, A., Kamiran, F., Ali, W., & Zhang, X. (2013). Controlling attribute effect in linear regression. *2013 IEEE 13th international conference on data mining*, 71–80.

Cohen, G. A. (2008). *Rescuing justice and equality*. Harvard University Press.

Collinson, P. (2017). EU's gender ruling on car insurance has made inequality worse. The Guardian. http://www.theguardian.com/money/blog/2017/jan/14/eu-gender-ruling-car-insurance-inequality-worse

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 797–806. https://doi.org/10.1145/3097983.3098095.

Custers, B., Calders, T., Schermer, B., & Zarsky, T. (Eds.). (2012). *Discrimination and privacy in the information society: Data mining and profiling in large databases* (2013th ed.). Springer.

Daniels, N. (2004). The functions of insurance and the fairness of genetic underwriting. In M. A. Rothstein (Ed.), *Genetics and life insurance: Medical underwriting and social policy (pp. 119–145)*. MIT Press.

Durnin, M., Hoy, M., & Ruse, M. (2012). Genetic testing and insurance: the complexity of adverse selection. *Ethical Perspectives, 19*(1), 123–154.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. https://doi.org/10.1145/2090236.2090255.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268. https://doi.org/10.1145/2783258.2783311.

Gajane, P. (2017). On formalizing fairness in prediction with machine learning. *ArXiv:1710.03184 [Cs, Stat]* http://arxiv.org/abs/1710.03184.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 3315–3323.

Heath, J. (2007). Reasonable restrictions on underwriting. In *Insurance ethics for a more ethical world* (Vol. 7, pp. 127–159). Emerald Group Publishing Limited.. https://doi.org/10.1016/S1529-2096(06)07007-6.

Hu, L., & Kohler-Hausmann, I. (2020). What's sex got to do with machine learning? *Proceedings of the 2020 conference on fairness, accountability, and transparency, 513*. https://doi.org/10.1145/3351095.3375674.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In P. A. Flach, T. De Bie, & N. Cristianini (Eds.), *Machine learning and knowledge discovery in databases* (Vol. 7524, pp. 35–50). Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-642-33486-3_3.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *8th innovations in theoretical computer science conference (ITCS 2017)* (Vol. 67, p. 43:1–43:23). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

Komiyama, J., Takeda, A., Honda, J., & Shimao, H. (2018). Nonconvex optimization for regression with fairness constraints. *International conference on machine learning*, 2737–2746 http://proceedings.mlr.press/v80/komiyama18a.html.

Lippert-Rasmussen, K. (2007). Nothing personal: On statistical discrimination*. *Journal of Political Philosophy, 15*(4), 385–403. https://doi.org/10.1111/j.1467-9760.2007.00285.x.

Lippert-Rasmussen, K. (2014). *Born free and equal? A philosophical inquiry into the nature of discrimination*. Oxford University Press.

Lipton, Z. C., Chouldechova, A., & McAuley, J. (2018). Does mitigating ML's impact disparity require treatment disparity? *Advances in Neural Information Processing Systems, 31* http://arxiv.org/abs/1711.07076.

MacAskill, W., & Ord, T. (2020). Why maximize expected choice-worthiness? *Noûs, 54*(2), 327–353. https://doi.org/10.1111/nous.12264.

Martin, K. E. (2018). Ethical implications and accountability of algorithms. *Journal of Business Ethics, forthcoming.* https://doi.org/10.1007/s10551-018-3921-3.

Menzies, P., & Beebee, H. (2020). Counterfactual theories of causation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Summer 2020)*. Metaphysics Research Lab: Stanford University https://plato.stanford.edu/archives/sum2020/entries/causation-counterfactual/.

Meyer, R. B. (2004). The insurer perspective. In M. A. Rothstein (Ed.), *Genetics and life insurance: medical underwriting and social policy (pp. 28–47)*. MIT Press.

O'Neill, M. (2006). Genetic information, life insurance, and social justice. *The Monist, 89*(4), 567–592. https://doi.org/10.5840/monist20068948.

Palmer, D. E. (2007). Insurance, risk assessment and fairness: an ethical analysis. In *Insurance ethics for a more ethical world* (Vol. 7, pp. 113–126). Emerald Group Publishing Limited. https://doi.org/10.1016/S1529-2096(06)07006-4.

Parfit, D. (2003). Equality and priority. In D. Matravers & G. E. Pike (Eds.), *Debates in contemporary political philosophy: An anthology* (pp. 115–132). Routledge.

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 560–568).

Perez, N. (2011). On compensation and return: Can the 'continuing injustice argument' for compensating for historical injustices justify compensation for such injustices or the return of property? *Journal of Applied Philosophy, 28*(2), 151–168. https://doi.org/10.1111/j.1468-5930.2011.00518.x.

Rawls, J. (1996). *Political liberalism* (Expanded ed.). Columbia University Press.

Rawls, J. (1999). *A theory of justice* (2nd ed.). Harvard University Press.

Scanlon, T. M. (1975). Preference and urgency. *The Journal of Philosophy, 72*(19), 655–669. JSTOR. https://doi.org/10.2307/2024630.

Thaysen, J. D., & Albertsen, A. (2017). When bad things happen to good people: luck egalitarianism and costly rescues. *Politics, Philosophy & Economics, 16*(1), 93–112. https://doi.org/10.1177/1470594X16666017.

Vigen, T. (n.d.). *15 insane things that correlate with each other*. http://tylervigen.com/spurious-correlations

Wilkie, D. (1997). Mutuality and solidarity: assessing risks and sharing losses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 352*(1357), 1039–1044.

Woodward, J. (2005). *Making things happen osps (1 edition)*. Oxford University Press.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web* (pp. 1171–1180).