The Fairness of Machine Learning in Insurance: New Rags for an Old Man?

Laurence Barry[1] and **Arthur Charpentier**[2]
PARI,[1] UQAM [2]
laurence.barry@mail.huji.ac.il,[1] arthur.charpentier@gmail.com[2]

Abstract

Since the beginning of their history, insurers have been known to use data to classify and price risks. As such, they were confronted early on with the problem of fairness and discrimination associated with data. This issue is becoming increasingly important with access to more granular and behavioural data, and is evolving to reflect current technologies and societal concerns. By looking into earlier debates on discrimination, we show that some algorithmic biases are a renewed version of older ones, while others seem to reverse the previous order. Paradoxically, while the insurance practice has not deeply changed nor are most of these biases new, the machine learning era still deeply shakes the conception of insurance fairness.

Introduction

Since the beginning of their history, insurers have been quantifying human phenomena, collecting and using data to classify and price risks. This practice is not trivial: insurance plays a major role in industrialized societies in opening or closing life opportunities (Baker and Simon 2002; Horan 2021). As such, insurers were confronted early on with the equity issues associated with data. Highlighting a bias means taking a critical stance on a calculation and shedding light on its political implications. This contestation is also culturally and historically situated.

As early as 1909, the Kansas insurance regulator thus drew the contours of fair rating practice: he defined a rating as "*not unfairly discriminatory*" if it treats similar people in the same way (Miller 2009; Frezal and Barry 2020). Based on this principle, certain parameters in insurance pricing were challenged during the 20th century, hence refining a typology of potential biases linked to classical data processing. The first section presents a historical perspective on insurance biases, showing how they were reformulated, clarified and reorganized with cultural and technological developments.

The second section describes the changes implied by the recent emergence of big data and new algorithms on insurance, and their induced biases. These debates actually revive and renew, with obvious points of continuity and rupture, older debates linked to discrimination in insurance. Conceptually however, these changes are shaking up insurance fairness: the 'individualization' of risks, rendered possible by these new technologies, is indeed now considered as fairer than their pooling. Finally, the last part discusses the ethical issues of big data for insurance, in comparison with various notions of algorithmic fairness.

Insurance and fairness: retrieving types of bias from historical debates

Pricing practice before machine learning

Insurance consists in the pooling of uncertainty: the contribution of the many makes it possible to compensate for the accidents of the few unluckiest. In its crudest form, the insurance risk premium is the mathematical expectation, on the group at stake, of the cost of the accident. Without competition between insurers, a single rate for all, as the average risk for the whole population, would do. Competition leads however to the threat of anti-selection: by segmenting the premium, insurer A can attract the best risks, thereby increasing his share at the expense of his competitors, who will make losses if they do not adopt a similar strategy. Segmentation therefore quickly becomes the rule of the game.

For a very long time, this segmentation consisted in the creation of supposedly homogeneous classes, on which the risk was estimated on average (Charpentier et al. 2015).

Before any calculation, the actuary's had to choose variables, a choice that projected homogeneity on the world, in two main ways: in the choice of what to ignore on the one hand (as what is not collected contains some heterogeneity that will not be seen); in the categorization of what is collected on the other hand, which again leads to crushing potential differences.

Gradually, the idea of a "perfect rate" emerged, in which the class would only include truly identical risks. Following Denuit and Charpentier (2004), and admitting that theta is the variable that would perfectly characterize the risk :

	Insured	Insurer
Loss	$\mathbb{E}[Y\|\Theta]$	$Y - \mathbb{E}[Y\|\Theta]$
Average Loss	$\mathbb{E}[Y]$	0
Variance	$\mathrm{Var}\left[\mathbb{E}[Y\|\Theta]\right]$	$\mathrm{Var}\left[Y - \mathbb{E}[Y\|\Theta]\right]$

Fig. 1: Risk split between insured and insurer

The variance on the portfolio is thus distributed between the policyholders who pay premiums proportional to their risk (captured by theta) and the insurer who carries the residual variance, unexplained by theta. In the 1980s, De Wit et Van Eeghen (1984) claimed that the increasing data collection and computational capabilities made it possible to refine the explained variance (and segmented premiums), thereby reducing the portion carried by the insurer.

However, the theta parameter, supposedly perfectly catching the risk, is never known. This uncertainty is the very basis of insurance: in a model for death benefits for instance, the probability of death can be estimated more precisely (some people have a 1 in 10,000 chance of dying and others a 1 in 1,000 chance), but it is impossible to predict *who* will die in a year. This fundamental residual uncertainty will irreducibly remain the insurer's responsibility, thus covered by the law of large numbers. Classification is then rethought as a means of approaching theta: the goal is no longer to simply counter anti-selection with ever finer classes, but also to have the unexplained variance converge towards a minimum. The actuarial practice does not change, even if its meaning does.

It is in this adjustment framework that the notion of bias appears: in the hypothesis that an exact calculation of the risk is possible, it occurs if the classification is imperfect and leads to the mispricing of certain groups, or persons, creating cross-subsidies between insureds (Walters 1981; De Pril and Dhaene 1996).

Historical Controversies on Insurance Pricing

From the 1960s onwards in the United States, the classification of risks was called into question in two main respects. First, in the context of the struggle for Black civil rights, the practice of 'red lining' or the exclusion of certain geographical areas from insured portfolios was criticized. Later, at the end of the 1970s, feminist movements tried to counter the use of gender in underwriting and pricing (Horan 2021). A detailed examination of the arguments allows to specify different aspects of what can be called "biased pricing," leading to a typology of pre-machine learning insurance biases that will serve our examination of machine learning in the next section.

One way to look at classification is to take it as a method for an ex-ante distribution of future costs, always more or less arbitrary. As such, the quantification work (of the statistician or actuary) is criticized because it is fed by an always socially constructed vision of the world (Desrosières 2008). Glenn (2000) points out that, like the Roman god Janus, an insurer's risk selection process has actually two faces: on the one side there is the face of numbers, actuarial tables, and statistics, that claims objectivity and rationality. But on the other side, there is the face of stories and subjective judgement. For Glenn, the actuary creates a myth in which decisions appear to be objective when in fact they are based on a great deal of subjectivity, prejudice, and stereotyping. These stereotypes come upstream the construction of the actuarial tables, in the stories that insurance technicians (actuaries and underwriters) tell each other, which lead them to collect one variable or another. Indeed, since data are collected via questionnaires, they are necessarily oriented by the pre-conception of risk held by the designer, itself socially constructed. These prejudices also appear downstream, in the story one tells to explain the results, therefore reinforcing existing social biases.

Race was among those factors. In life insurance for instance, Bouk (2015, 34) describes how, at the end of the 19th century, insurers charged the same premium to everyone but paid claims differentially according to skin colour (see also Heen 2009). Several states then passed anti-discrimination laws. Thus, in the summer of 1884, the state of Massachusetts enacted a law prohibiting "any distinction or discrimination between white persons and colored persons wholly or partially of African descent, as to the premiums or rates charged for policies upon the lives of such persons" (quoted in Wiggins 2013, 68). To counter the law, Frederick L. Hoffman, supported by Prudential Life Insurance, published a book in 1896 demonstrating the higher mortality of Black Americans (Heen 2009, 377; Bouk 2015, 49–52). Insuring them at the same rate as Whites would be statistically inequitable, he argued; not insuring them was therefore the only way to comply with the law, which in fact made Black Americans uninsurable.

Moreover, in the controversies surrounding classification, Horan (2021, 170–71) shows that pricing parameters also evolve in response to regulatory, political or social constraints: "the categories insurance companies used to create risk classifications throughout the twentieth century

reflected changing political trends and social values, and not simply objective realities." Alternative classifications can be equally effective, highlighting the room for manoeuvre, but also the arbitrariness of the decisions left to practitioners:

> Insurers can rate risks in many different ways depending on the stories they tell about which characteristics are important and which are not (...) The fact that the selection of risk factors is subjective and contingent upon narratives of risk and responsibility has in the past played a far larger role than whether or not someone with a wood stove is charged higher premiums (Glenn 2003, 135).

While Glenn insists on the subjective selection of factors, one can argue that this selection, that aspires to scientific exactitude and objectivity, is actually oriented by what is recognized as true and/or acceptable in a given time and place.

For Schauer (2003) two types of stereotypes should be distinguished. Some generalizations are completely unfounded: generalizations based on a person's astrological sign, for example, are pure prejudice. But others have a statistical basis, when the probability of having a character y knowing x is significantly different from the case where nothing is known. From this perspective, the use of the male/female parameter remains legitimate because it is statistically significant for estimating a probability of death or of a car accident. Any classification based on variables that are effectively correlated with the risk would then be legitimate.

Works (1977) warns, however, against "proxy variables," as opposed to "true variables" of risk. The latter would be more difficult to obtain, and thus replaced by simple approximations - leaving again the door open to all kinds of biases in pricing and underwriting. In the 1980s class actions against the use of the gender parameter, this was the approach taken by the plaintiffs. Their main argument was that the observed correlation between the cost of motor insurance claims and the driver's gender was due to the lower mileage of women; mileage is the causal variable, and therefore legitimate, and not gender, which is only a biased approximation of the latter (Horan 2021). The underlying hypothesis here is that there are "true variables" of risk which explain accidents in a causal manner, all the others being invalid.

The problem with this type of argument is that the existence of a direct causality is very difficult to establish, and therefore it is more of a judgement, again socially constructed, than a real scientific proof: for example, when Hoffman at the end of the 19th century highlighted a correlation between life expectancy and skin colour, he deduced the existence of an innate causality linked to the Black race that made it more risky, whereas others would have sought environmental and social causes explaining the greater mortality of Blacks (Heen 2009, 377). Hence causal relations too might simply be the latest version of a socially acceptable explanation.

Moreover, the use of proxy variables can also come from an intention to avoid regulation. For instance, once racial parameters became forbidden, insurers and other financial institutions started delineating areas according to their racial occupation. Ethnicity can indeed be inferred with a fair degree of accuracy from the location of potential policyholders. A survey commissioned by the federal government in the 1960s thus revealed that many financial institutions refused to serve predominantly African-American geographic areas (Austin 1983; Horan 2021), and that this systematic practice of "red-lining" had led to a deterioration of services and infrastructure in certain cities. This discrimination seems to persist to this day: Larson et al. (2017) conducted an analysis by zip code for major insurance companies across the US, showing that the average premium is 10% higher in auto liability for zip codes associated with minority populations.

Besides, even if a causality has been recognized as valid, it remains open to political controversies. Simon (1988, 795-6) for instance argues that causality or correlation ultimately do not matter when it comes to fighting blatant social discriminations: on this basis, the use of the discriminating parameter contributes to naturalizing the difference in (social) treatment and thus to anchoring the discriminatory reality. This happens for instance with credit-based insurance scoring, that has given way to intensive debates in the United States, starting in the 1990s up to this day. While credit-based scores properly predict insurance losses, the National Association of Insurance Commissioners (NAIC) wanted to better understand *why* that was the case. In her study, Kiviat (2019) shows how competing causal theories were proposed in the numerous hearings: on one side, insurers defended the idea that higher credit scores reflect careful behavior; on the other, public representatives demonstrated that scores predicted the propensity to file a claim in case of an accident, but not the likelihood of having an accident. Hence the "real" cause was the economic status of the insured, which made credit-based insurance rating an unfair practice: policymakers were indeed "less willing to accept that consumers might deserve to pay more by virtue of their low earning power" (Kiviat 2019, 1146).

In this understanding, a biased model reinforces existing discrimination. In a classical statistical set-up, the solution is to prohibit the use of the incriminated variable. Today, the list of such protected factors varies greatly between countries. In Europe for instance, it currently includes religious beliefs, sexual orientation, trade union involvement, ethnicity, medical status, criminal convictions and offences, biometric data, genetic information and, most recently, gender.

Interestingly, in the European Union 2011 decision that prohibits the use of gender in insurance prices, the judge further distinguishes between two types of causal variables, pointing to what could be considered a fair classification, once non-significant and non-causal variables are eliminated: "Like race and ethnic origin, sex is also an inseparable characteristic of the insured person over which *he or she has no influence*" (CURIA 2010, emphasis added). This distinction refers to what is known in the literature as "brute and option luck" (Dworkin 1981): some hazards are linked to personal choices (*option luck*) and are thus voluntary; others are caused by elements over which the individual has no control (*brute luck*). The judge thus implies that while the former can be used for pricing, the latter must be taken care of by the community (and therefore protected and eliminated from pricing).

This distinction is however even more difficult to establish than the causal inference. In the case of natural catastrophes for instance, some have argued that to live in a risky area is a choice, hence the pricing according to geography was legitimate; others however insist that some less well-off populations have no other choice but to live in cheaper yet riskier areas (Charpentier et al. 2021).

The critics of classical classifications are thus distributed along three types of biases:

- Type 1 biases are linked to classes that do not reflect the reality of the risk, either by mistake or due to pure prejudice. In statistical terms, the model is bluntly wrong or reflects spurious correlations. In social ones, it comes to perpetuate existing discriminations in the form of what Barocas and Selbst (2016, 694) call disparate treatment: it "comprises two different strains of discrimination: (1) formal disparate treatment of similarly situated people and (2) intent to discriminate."

- Type 2 biases are linked to classes that reflect a proven statistical reality (a correlation with risk), but the variables are known to be non-causal. This makes them suspect of bias and arbitrary choice, in the same way as type 1 are. This is the case for instance with redlining, where the variable is used as a proxy of race to discriminate against minorities.

- Type 3 biases are linked to classes that reflect a statistical reality, that is taken to be causal. However, for some reason, the classification is deemed unacceptable on other grounds. Although utterly correct, it is intrinsically harmful because it reproduces and anchors in reality a situation that must be fought against. Type 3 biases also include cases where the causal variable does not capture an *existing* social discrimination, but its use for pricing would *create* one. Differentiating risks based on genetic data in health insurance would be such an example. Type 3 biases thus also establish a distinction between causal elements that are voluntary or non-voluntary, the pricing based on the latter being considered unfair.

Interestingly, the distinction between type 2 and 3 biases is as difficult as proving causality. The distinction thus comes from whether the explanation provided by the model makes (causal) sense in a given time and place. But this further triggers questions, such as - when is a correlation admitted as a cause, and when is it not? And, once recognized as a cause, when is it perceived as acceptable, and why? Trying to answer these questions would go far beyond the scope of this paper. However, we will try to show in the next section how these become crucial with machine learning.

From Classification to Machine Learning

From the 2000s onwards, big data and new algorithms have implied a displacement of tasks between humans and machines. Measuring the impact on insurance is however maybe more difficult than in other areas. On the one hand, like any other organization, insurers are pushed to change their practices to incorporate the new sources of data that have become available, the increased computing capacity, and the new algorithms. This has obvious consequences on the risk classification process. On the other hand, the new techniques often appear to be the continuation of this almost age-old segmentation practice (Swedloff 2014). From this perspective, we will show below that the new algorithms are only renewing existing questionings on discrimination and biases.

The Datafication of Risk

According to Ewald (2011), the new era means that the quantification of hazards, and the way to protect society against them, is changing: we would be moving from risk, where hazards were dealt with by classifying and mutualizing hazard into homogeneous classes, to "data." The "datafication" of risk (Mayer-Schönberger and Cukier 2014) then means that hazards are not calculated via the classification process, but by extensive use of individual data. However, some studies show that, to date at least, pricing models have not changed significantly, nor have new products emerged beyond pilot experiments (Barry and Charpentier 2020; François and Voldoire 2022). The study below is therefore more of an analysis of what the new models *make possible*, even if the change in insurance has not (yet?) been observed in practice, leading Meyers (2018) to speak of "not-yet markets."

The change appears first as conceptual. In her opinion concerning the prohibition of the use of gender in insurance pricing, the European judge thus invokes the imprecision of statistics, when what is at stake, in her view, is the individual risk (CURIA 2010). She moreover seems to encourage pricing based on the behavioural data recently made available by sensors (Meyers 2018), and is thus in line with the trend that associates risk with lifestyle rather than statistical

classes (Rebert and Van Hoyweghen 2015). The utopic claim that today's algorithms would be capable of personalizing decisions, whereas their ancestors were only grossly working on averages (Moor and Lury 2018; Lury and Day 2019) seems therefore to be in the process of being transposed to insurance.

This conceptual shift is driven by the emergence of big data. In sharp contrast with the era of questionnaires, today data are indeed obtained via sensors, connected objects or coming from online actions; they are thus *natively digital* - all sources that do not require human intervention. Moreover, these data are more often than not behavioural, and almost continuous (Barry and Charpentier 2020). The second major change is that computational capabilities are now far greater than with the previous generation of computers. This processing of much larger databases is changing the environment of data analysis.

Finally, machine learning allows for the automation of (some?) of the tasks, particularly that of choosing significant variables, which increases the number of variables that can be considered. Models thus become more complex, without necessarily changing in nature. A conceptual leap occurs, however, with deep learning algorithms (taken here as a category of machine learning). LeCun, Bengio, et Hinton (2015, 436) indeed characterize deep learning by its ability to infer potential relationships between variables, where these were previously imposed on the data by the analyst.

Put in perspective with the previous section, machine learning seems to remove type 1 and 2 biases that resulted from the actuary's prejudices and stereotypes in his choice and coding of variables. The recent access to behavioural data also seems to meet the need to distinguish between variables describing conscious choices of the insured (his behaviour) and those relating to intrinsic characteristics over which he has no control, hence lifting part of the difficulties linked to type 3 biases and brute versus option luck. Would big data and machine learning thus allow to lift the discriminations associated with classification? We'll show below that nothing is less certain, even if the claim that they could is recurringly heard.

The biases of machine learning in insurance

The biases delineated in the first section were linked respectively to social perceptions of the world that influenced the statistician, to correlations that were not causation, and to causal relations that should be corrected rather than reproduced in the models. Interestingly, each of these biases find a new expression and answer with machine learning.

Risks' "Ground Truth"

"Ground truth" is an important and interesting concept of machine learning. At first hand, the meaning seems clear – it is the physical reality that the algorithm comes to predict. Wikipedia's definition is "the knowledge of the truth concerning a specific question. It is the ideal expected result." It refers however to the "ground truth" *as seen by the algorithm*, that is as mediated by the data taken to describe the physical reality at stake (Grosman and Reigeluth 2019; Jaton 2021). But to what extent is this representation accurate? To what extent does "ground truth" actually reflect reality?

Insurance companies are increasingly relying on data from external sources for modelling their risks: most frequently, these are online actions, be they invoices, transactions, emails, photos, click streams, logs, search queries, medical records, etc. (Charpentier 2021). In the somewhat mythological utopia of big data, this data would finally have become exhaustive, allowing a richer (full?) account of reality: without the compromise of sampling, without the constraints of volume (Mayer-Schönberger and Cukier 2014) and without the reduction of reality due to quantification problems (Desrosières 2008).

One key issue however is that these data are observational, leading -paradoxically for the myth of exhaustivity, to multiple biases (Rosenbaum 2017; Charpentier 2021). The distinction between observational and experience data was first established by Ronald Fisher in 1935 (Fisher 1971) when devising randomized experiments, to solve type 2 biases and separate between causal and non-causal relationships. The method consists in setting up two similar (if not identical) groups with which to measure the effect of a variable (a treatment in medicine, for example). The proximity is measured by a series of co-variables (age, gender, economic status, etc), that all might affect the result and are hence "under control." One of the groups, known as the control group, is indeed not subjected to the treatment, allowing to observe the effect of treatment on the second group, *all other things being equal* (Rosenbaum 2017; Leigh 2018).

Resulting, in most cases, from the behaviour of people independently of any ad hoc experiment, big data are observational hence cannot, without further qualification, prove causation. An example of such a bias in data science is the one given by Caruana et al. (2015): researchers in the 1990s trained a deep neural network to classify patients as low or high risk for pneumonia, in order to limit hospital admissions. The model was extremely accurate on the training data, but the results were counter-intuitive on asthma patients, who were classified as low risk. A closer examination showed that asthma patients were treated much more quickly, precisely because of their very high mortality risk in case of pneumonia. Their low risk was therefore the result of differential treatment (the two populations were not identical) which the algorithm did not take into account. This example shows that one cannot fully make sense of the data without knowing the stories behind them, that is *the process that led to their collection*. As Pearl and Mackenzie (2018, 206) put it, "we must look beyond the data to the data generating process."

Besides, big data magnify type 1 sample biases of the classical statistics era, such as Hoffman's choosing specific or partial populations to compare mortality rates between Blacks and Whites (Bouk 2015). With big data, the filtering is no longer the responsibility of the statistician who builds his database, yet it exists all the same (boyd and Crawford 2012). For Barocas et Selbst (2016), we have moved from disparate treatment, that is Hoffman consciously building *ad hoc* data to prove a point, to disparate impact. This happens non intentionally when populations at the margins of the formal economy and/or online activities are under-represented in the data used to train the algorithms. The non-intentionality and potential transparence of the problem to the expert aggravates the discrimination issue, the more so as models have become more complex.

Another known big data sample bias is related to self-selection (Charpentier 2021). This situation is increasingly found in European public administrations' data bases for instance, where until very recently data were stored automatically. With the General Data Protection Regulation (GDPR) - whose main aim is the protection of personal data, it is now possible for those who request it to have their data deleted. This concept of *opting-out* can strongly bias the data retained. In all these examples, the actuary, who has not himself constructed the databases to which he now has access, sometimes has difficulty understanding their limitations.

Interpretability vs. Accuracy: the Opaque Effectiveness of New Algorithms

The new algorithms' often decried opacity is usually weighed against their increased accuracy compared to classical models (Breiman 2001). In 2017, during one of the first debates at the NeurIPS conference, it was pointed out that "if we wish to make AI systems deployed on self-driving cars safe, straightforward black-box models will not suffice, as we need methods of understanding their rare but costly mistakes." At the conference, Yann LeCun claimed that, when presented with two models (one perfectly interpretable and 90% accurate, the other a black-box with a higher accuracy of 99%), people always chose the more accurate model. He concludes that "people don't really care about interpretability but just want some sort of reassurance from the working model" (*The Great AI Debate: Interpretability is necessary for machine learning* 2017). Interpretability is not important once one is convinced that the model works well under the conditions in which it is supposed to work. This claim completely displaces the understanding of type 2 biases (the attempt to separate correlation from causation) in the big data era.

For Napoletani, Panza, et Struppa (2011, 3), deep learning opens the way to a new kind of science that has become "agnostic" to causes. As such, in a now famous article Anderson (2008) speaks of the "end of theories" to characterize this new approach, that leads to the renunciation of the need to highlight causal links between variables (and therefore to explain the phenomenon). From the perspective of this paper, this would mean giving up the explicitation of the relationships between variables, either causal or correlated: the machine produces a score, sufficiently precise to justify giving up an interpretation.

There are a few (still rare) examples of this black-box approach in insurance. For example, image recognition algorithms can infer risk factors. Kita-Wojciechowska et Kidziński (2019) thus propose to predict the frequency of car accidents from satellite images of the driver's place of residence; or Shikhare (2021) calculates a health score based on a photo portrait of the person.

The efficacity of black-box algorithms cannot however lift the discrimination issues, to the contrary. Paradoxically, while in the previous era, the stories told by statisticians were incriminated as leading to socially constructed biases, it is now the impossibility of telling (causal) stories that poses a problem. As Kiviat (2019, 1152) puts it "what gets wiped away along with storytelling is an ability to appreciate how bad luck or the inequities of history can set events in motion and cause people to show up in the data in particular ways."

Correcting social discrimination: protecting data in a big data environment

Moreover, while the machine learning process is supposed to be agnostic to social constructs and biases, researchers have now proven that, on the contrary, prejudices, stereotypes and other discriminations are found in the data themselves, and therefore well upstream of the statisticians' judgement: beyond the sample bias mentioned above, it is really the nature of the data that is at issue (Caliskan et al. 2017).

Besides, whereas in classical models one could hope to correct biases by prohibiting the use of so-called protected variables, the collinearity of these variables with other, facially neutral ones in big data makes this "protection" illusory: "thus, a data mining model with a large number of variables will determine the extent to which membership in a protected class is relevant to the sought-after trait *whether or not that information is an input*" (Barocas et Selbst 2016, emphasis added)

For Prince et Schwarcz (2019) the proxy discrimination, already mentioned for classical models, is magnified by the new algorithms. Whereas it was intentional in the past (since a human decision presided over the choice of variables), proxy discrimination becomes unintentional. This phenomenon is unavoidable, especially when a variable directly related to the phenomenon (a causal variable) is absent from the data:

> To illustrate, an AI deprived of information about a person's genetic test results or obvious proxies for this information (like family history) will use other

> information-ranging from TV viewing habits to spending habits to geolocational data-to proxy for the directly predictive information contained within the genetic test results (Prince and Schwarcz 2019, 1274).

In this specific example, the algorithm might create a health risk factor based on the viewing of television programs! To combat this phenomenon, Williams, Brooks, et Shmargad (2018) show that the collection and use of protected variables should not be prohibited, but rather used as a means of monitoring non-discrimination instead. But this is not easily done either, as will be shown below.

Insurance Fairness in the Age of Machine Learning: Collective or Individual?

According to Thiery et Schoubroeck (2006), lawyers and actuaries have fundamentally different conceptions of equity. Legally, the right to equal treatment is granted to a person as an individual. But this view is fundamentally opposed to actuarial fairness, which historically builds on an analysis of risks and calculation of premiums in collective terms (Ewald 2011). For instance, in the 1983 Norris decision on a class action against the use of gender for retirement benefits, the judge maintained that a statistically valid classification (whether causal or correlational) does not make it a legitimate classification. In fact, no classification can be, since "even a true generalization about class cannot justify class-based treatment. An individual woman may not be paid lower monthly benefits simply because *women as a class* live longer than men" (quoted in Horan 2021, 187). For the individual on whom it is imposed, classification always results from a "statistical bias" (Binns 2018), that is an arbitrary inference from the group to the individual.

For Simon (1988) and Horan (2021), the adoption of this individual point of view by the judge in the Norris case has contributed to reinforcing the erasure of the solidarity principle which is at the heart of insurance practice (Lehtonen and Liukko 2011). Once the existence of an individual risk has been accepted, first approached through classification then through learning algorithms, pricing becomes a mathematical exercise in optimizing and minimizing the insurer's variance. Big data algorithmic "personalization" is translated into risk "individualization" in insurance (Barry and Charpentier 2020). Thus the distinction between insurance-collective versus legal-individual kinds of fairness tends to disappear (Barry 2020), giving to the "statistical discrimination" outlined by the judge in the Norris case a renewed importance.

Indeed, even if actuaries have not fundamentally changed their practice, the notion of insurance fairness is changing. Meyers and Van Hoyweghen (2018) thus show on a telematic product that risk is no longer presented as a pooled uncertainty, but as an individual choice. Individual behaviour should determine the premium, not aggregate demographics. Fairness in this case means adjusting the premium to individual behaviour, so that everyone pays according to "their" risk (Meyers and Van Hoyweghen 2018). For Fourcade and Healy (2017, 24), algorithmic scores rely on moral economy of *deserve*, where each is responsible for his acts and their consequences.

This individualization, if it takes place, triggers its own fairness and discrimination issues. First, since it would lead to more disparate pricing, individuals found as "the riskiest" by the algorithm are likely to get unaffordable premiums, hence excluding them from the insured community (Charpentier et al. 2020). Besides, research in algorithmic fairness, that is emerging as a new discipline (Kusner and Loftus 2020), shows that the tension between individual and collective viewpoints, at the heart of current questionings on the individualization of risk, interestingly finds a new forum in this literature.

Aggregate indicators of the fairness of an algorithm might come from its (mathematical) accuracy. It is generally measured by means of a confusion matrix, which allows to observe errors by type - false negatives and false positives. But simultaneously minimizing these errors is not possible, or even desirable, for several reasons. Indeed, false positives and false negatives are not comparable from an ethical point of view: the conviction of an innocent person does not have the same "value" as the release of a guilty person. Thus, depending on the context, it will be necessary to choose to minimize one or the other form of error.

Things become even more complicated when protected variables are taken into account. Pessach et Shmueli (2020) then distinguish between collective and individual fairness indicators. Collective ones aim to ensure parity between protected and non-protected groups. One can thus check that the (exact or positive) prediction frequencies, or that the false positive and false negative rates, calculated separately for each group are close. This was not the case, for example, for the *Correctional Offender Management Profiling for Alternative Sanctions* algorithm (COMPAS), which had a much higher false positive rate (falsely classified as recidivists) for Blacks, and a higher false negative rate (falsely classified as non-recidivists) for Whites, with equal accuracies on both groups (Kleinberg et al. 2016). Individual indicators, on the other hand, aim to ensure that similar individuals obtain a similar score. Kusner et Loftus (2020) thus define "counterfactual" equity, which consists of comparing the scores of two identical observations in which only the protected variable takes a different value. All authors agree that these different indicators cannot be optimized simultaneously, leading to necessary context-dependent trade-offs (Kleinberg et al. 2016; Pessach and Shmueli 2020).

The ban on the male/female parameter by the European directive exemplifies these dilemmas in insurance: either the variable is ignored, but then if a statistical difference exists

it will be captured through other variables, that are collinear with the banned parameter. Consequently, the average for men and women will remain different. Or, on the contrary, this variable is used to maintain identical averages, but then, all other things being equal, the rates will vary with the sex of the person. It will never be possible to maintain parity between the groups and ensure counterfactual equity. For Charpentier (2021, 148), prohibiting the use of the protected variable is counterproductive because "in most realistic cases, not only does the removal of the sensitive variable not make the regression models fair, but on the contrary, such a strategy is likely to amplify the discrimination."

Conclusion

Insurance fairness is a dynamic concept, which depends on historical, cultural and technical contexts. At the height of the industrial era, the veil of ignorance on individual hazards and the idea of the equal fate of all in front of adversity were used to justify very broad and solidary coverage. This conception of fairness was criticized by the most conservatives since it was seen as an incentive to license. During the 20th century, with the growing capacities of data gathering and calculation, segmented models were adopted, which based insurance on the classification of risks into homogeneous groups of people. From the 1980s onwards, controversies arose over the use of this or that variable, that frame the current criticism of machine learning's biases and discriminations. The examination of this history allowed us to identify in traditional classification practices a few main families of bias, that could then be tracked in machine-learning algorithms.

The first claim is that when the parameters have no connection with the phenomenon to be studied, their (manual) choice only reflect the prejudices of the statistician. This is the case of skin colour in the United States in life insurance products at the end of the 19th century. This type of bias disappears in principle with big data: being natively digital, they bypass the quantification work of the previous period. However, over the last twenty years, researchers have shown that data absorb social discrimination and prejudices. Blind use of machine learning would then lead to the reproduction of these biases in algorithmic decisions.

It was soon found that discarding statistically irrelevant variables is not enough. In the 1960s and 1980s, the use of correlated but not causal variables, that is magnified with the new algorithms, became the main source of complaint: the male/female parameter, or the credit score among others have provoked controversies of this kind, some of which continue to this day. The requirement of proven causality, already inoperable in the statistical era, has been totally abandoned in machine-learning algorithms. Some indeed say that big data marks the advent of a new *episteme*, where finding correlations and patterns in the input data, without even making these links explicit, would be the core of the new science. But this displaces the criticism towards the algorithms' opacity, that despite greater precision create their own implicit biases.

Moreover, some causal variables reflect hazard that were not chosen by the ones that endure them. In this case, insurance is seen as a way *not to* reflect the risk but, by eliminating the variable from the models, to have it borne by the entire insured population: the use of genetic data in health insurance, for example, is prohibited today in most countries. But the solution of eliminating protected variables, effective in traditional models, is much more difficult to implement with big data and machine learning, respectively because protected variables are captured via their collinearity with others, and the opacity of the algorithms makes it more complex to highlight these discriminations.

More fundamentally, a legalistic critique opposed fairness as an *individual* right to the collective insurance approach. In this strand of thought, the necessarily arbitrary reduction of an individual to the data of a class was seen as a *statistical* bias. Big and behavioral data, that sign the advent of personalization and the potential individualization of risk, supposedly solve this statistical bias by promoting in insurance the fairness of deserve, where each pay for the risks he chooses to take. But here again, the promise is not kept: researchers in algorithmic fairness indeed highlight the impossibility of optimizing algorithms on multiple criteria, none of which can be a priori preferred to another. In the insurance context, individual fairness also threatens to lead to increasingly differentiated rates, therefore making protection unaffordable for those classified as very risky.

This history also interestingly shows that once segmentation is adopted as a principle, the legitimacy of an insurance model relies on its being both "causal," in the sense of explaining risk, and fair. Yet both causality and fairness are strongly contingent upon the narratives seen as valid in a given time and place. Besides, current state of the art algorithms cannot produce models that satisfactorily provide (causal) explanations. Should insurance then stick to the good old pricing tables, for which all the parameters are explicit, known in advance and therefore open to challenge? This does not warrant fairness, but *contestability*: just as any scientific theory must be falsifiable, a pricing system should be transparent in order to be contestable.

Acknowledgement

This study was supported by the Chaire PARI - project « Evaluation des Risques et Technologies du Big Data :

Outils et Conséquences », Fondation Institut Europlace de Finance.

References

Anderson, C. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*.

Austin, R. 1983. The Insurance Classification Controversy. *University of Pennsylvania Law Review* 131: 517–582. doi:10.2307/3311844.

Baker, T., and J. Simon. 2002. Embracing Risk. In *Embracing Risk: The Changing Culture of Insurance and Responsibility*, 1–25. University of Chicago Press.

Barocas, S., and A. D. Selbst. 2016. Big Data's Disparate Impact Essay. *California Law Review* 104: 671–732.

Barry, L. 2020. Insurance, Big Data and Changing Conceptions of Fairness. *European Journal of Sociology / Archives Européennes de Sociologie* 61. Cambridge University Press: 159–184. doi:10.1017/S0003975620000089.

Barry, L., and A. Charpentier. 2020. Personalization as a promise: Can Big Data change the practice of insurance? *Big Data & Society*.

Binns, R. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Conference on Fairness, Accountability and Transparency*, 149–159. PMLR.

Bouk, D. 2015. *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual*. Chicago ; London: University Of Chicago Press.

boyd, dana, and K. Crawford. 2012. Critical Questions for Big Data. *Information, Communication and Society* 15: 662–679. doi:10.1080/1369118X.2012.678878.

Breiman, L. 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science* 16: 199–231. doi:10.1214/ss/1009213726.

Caliskan, A., J. J. Bryson, and A. Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*. American Association for the Advancement of Science. World. doi:10.1126/science.aal4230.

Caruana, R., Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. KDD '15. New York, NY, USA: Association for Computing Machinery. doi:10.1145/2783258.2788613.

Charpentier, A. 2021. *Assurance : Bais, Discrimination & Équité*. unpublished manuscript.

Charpentier, A., M. M. Denuit, and R. Elie. 2015. Segmentation et Mutualisation, les deux faces d'une même pièce. *Risques*: 19–23.

Charpentier, A., L. Barry, and E. Gallic. 2020. Quel avenir pour les probabilités prédictives en assurance ? *Annales des Mines - Realites industrielles* 2020: 74–77.

Charpentier, A., L. Barry, and M. R. James. 2021. Insurance against natural catastrophes: balancing actuarial fairness and social solidarity. *The Geneva Papers on Risk and Insurance - Issues and Practice*: 1–29. doi:10.1057/s41288-021-00233-7.

CURIA. 2010. *Test-Achats Conclusions de l'Avocat General*. Court of Justice of the European Union.

De Pril, N., and J. Dhaene. 1996. *Segmentering in verzekeringen.* Leuven: KUL. Departement toegepaste economische wetenschappen.

De Wit, G. W., and J. Van Eeghen. 1984. Rate Making and Society's Sense of Fairness. *ASTIN Bulletin*: 151–164.

Denuit, M., and A. Charpentier. 2004. *Mathématiques de l'assurance non-vie: Principes fondamentaux de théorie du risque*. ECONOMICA edition. Paris: ECONOMICA.

Desrosières, A. 2008. *L'argument statistique. I, Pour une sociologie historique de la quantification*. Paris: Presses de l'école des Mines.

Dworkin, R. 1981. What is Equality? Part 2: Equality of Resources. *Philosophy & Public Affairs* 10. Wiley: 283–345.

Ewald, F. 2011. Omnes et Singulatim. After Risk. *Carceral Notebooks* 7: 77–107.

Fisher, R. A. 1971. *The Design of Experiments*. New York: Macmillan Pub Co.

Fourcade, M., and K. Healy. 2017. Seeing like a Market. *Socio-economic review* 15: 9–29.

François, P., and T. Voldoire. 2022. *The revolution that did not happen. Telematics and car insurance in the 2010s*. Working Paper 27. Paris: Chaire PARI.

Frezal, S., and L. Barry. 2020. Fairness in Uncertainty: Some Limits and Misinterpretations of Actuarial Fairness. *Journal of Business Ethics* 167: 127–136. doi:10.1007/s10551-019-04171-2.

Glenn, B. J. 2000. The Shifting Rhetoric of Insurance Denial. *Law & Society Review* 34. [Wiley, Law and Society Association]: 779–808. doi:10.2307/3115143.

Glenn, B. J. 2003. Postmodernism: The Basis of Insurance. *Risk Management & Insurance Review* 6. Wiley-Blackwell: 131–143. doi:10.1046/J.1098-1616.2003.028.x.

Grosman, J., and T. Reigeluth. 2019. Perspectives on algorithmic normativities: engineers, objects, activities. *Big Data & Society* 6. SAGE Publications Ltd: 2053951719858742. doi:10.1177/2053951719858742.

Heen, M. 2009. Ending Jim Crow Life Insurance Rates. *Northwestern Journal of Law & Social Policy* 4: 360–399.

Horan, C. D. 2021. *Insurance Era: Risk, Governance, and the Privatization of Security in Postwar America*. First edition. Chicago ; London: University of Chicago Press.

Jaton, F. 2021. Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society* 8. SAGE Publications Ltd: 20539517211013570. doi:10.1177/20539517211013569.

Kita-Wojciechowska, K., and L. Kidziński. 2019. Google Street View image predicts car accident risk. *Central European Economic Journal* 6. Wydawnictwo Uniwersytetu Warszawskiego: 152–163.

Kiviat, B. 2019. The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores. *American Sociological Review* 84. SAGE Publications Inc: 1134–1158. doi:10.1177/0003122419884917.

Kleinberg, J., S. Mullainathan, and M. Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores.

Kusner, M. J., and J. R. Loftus. 2020. The long road to fairer algorithms. *Nature* 578: 34–36. doi:10.1038/d41586-020-00274-3.

Larson, J., J. Angwin, L. Kirchner, and S. Mattu. 2017. How We Examined Racial Discrimination in Auto Insurance Prices. *ProPublica*.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521: 436–444. doi:10.1038/nature14539.

Lehtonen, T.-K., and J. Liukko. 2011. The Forms and Limits of Insurance Solidarity. *Journal of Business Ethics* 103: 33–44. doi:10.1007/s10551-012-1221-x.

Leigh, A. 2018. *Randomistas: How Radical Researchers Are Changing Our World*. New Haven: Yale University Press.

Lury, C., and S. Day. 2019. Algorithmic Personalization as a Mode of Individuation. *Theory, Culture & Society* 36: 17–37. doi:10.1177/0263276418818888.

Mayer-Schönberger, V., and K. Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Reprint edition. Boston: Eamon Dolan/Mariner Books.

Meyers, G. 2018. Behaviour-based personalisation in health insurance: a sociology of a not-yet market. PhD Thesis, KU Leuven.

Meyers, G., and I. Van Hoyweghen. 2018. Enacting Actuarial Fairness in Insurance: From Fair Discrimination to Behaviour-based Fairness. *Science as Culture* 27: 413–438. doi:10.1080/09505431.2017.1398223.

Miller, M. J. 2009. Disparate Impact and Unfairly Discriminatory Insurance Rates. *Casualty Actuarial Society E-Forum*.

Moor, L., and C. Lury. 2018. Price and the person: markets, discrimination, and personhood. *Journal of Cultural Economy* 11: 501–513. doi:10.1080/17530350.2018.1481878.

Napoletani, D., M. Panza, and D. C. Struppa. 2011. Agnostic Science. Towards a Philosophy of Data Analysis. *Foundations of Science* 16: 1–20. doi:10.1007/s10699-010-9186-7.

Pearl, J., and D. Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. 1st edition. New York: Basic Books.

Pessach, D., and E. Shmueli. 2020. Algorithmic Fairness.

Prince, A. E. R., and D. Schwarcz. 2019. Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review* 105: 1257.

Rebert, L., and I. Van Hoyweghen. 2015. The right to underwrite gender. The Goods & Services Directive and the politics of insurance pricing. *Tijdschrift voor Genderstudies* 18. Uitgeverij Aksant: 413–431.

Rosenbaum, P. 2017. *Observation and Experiment: An Introduction to Causal Inference*. *Observation and Experiment*. Harvard University Press. doi:10.4159/9780674982697.

Schauer, F. 2003. *Profiles, Probabilities, and Stereotypes*. Harvard University Press. doi:10.2307/j.ctvjz82xm.

Shikhare, S. 2021. AI Enabled Next Generation LTC and Life Insurance Underwiring Using Facial Score Model. In *Insurance Data Science conference 2021*, 19. London.

Simon, J. 1988. The Ideological Effects of Actuarial Practices. *Law Social Review* 22: 771–800.

Swedloff, R. 2014. Risk Classification Big Data (R)Evolution. *Connecticut Insurance Law Journal* 21: 339–373.

The Great AI Debate: Interpretability is necessary for machine learning. 2017. Neurips 2017.

Thiery, Y., and C. V. Schoubroeck. 2006. Fairness and Equality in Insurance Classification. *The Geneva Papers on Risk and Insurance - Issues and Practice* 31: 190–211. doi:10.1057/palgrave.gpp.2510078.

Walters, M. A. 1981. Risk Classification Standards. *Proceedings of the Casualty Actuarial Society* 68: 1–23.

Wiggins, B. A. 2013. Managing risk, managing race: racialized actuarial science in the United States, 1881-1948. Minnesota.

Williams, B. A., C. F. Brooks, and Y. Shmargad. 2018. How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy* 8. Penn State University Press: 78–115. doi:10.5325/jinfopoli.8.2018.0078.

Works, R. 1977. Whatever's FAIR—Adequacy, Equity, and the Underwriting Prerogative in Property Insurance Markets. *Nebraska Law Review* 56: 445–464.