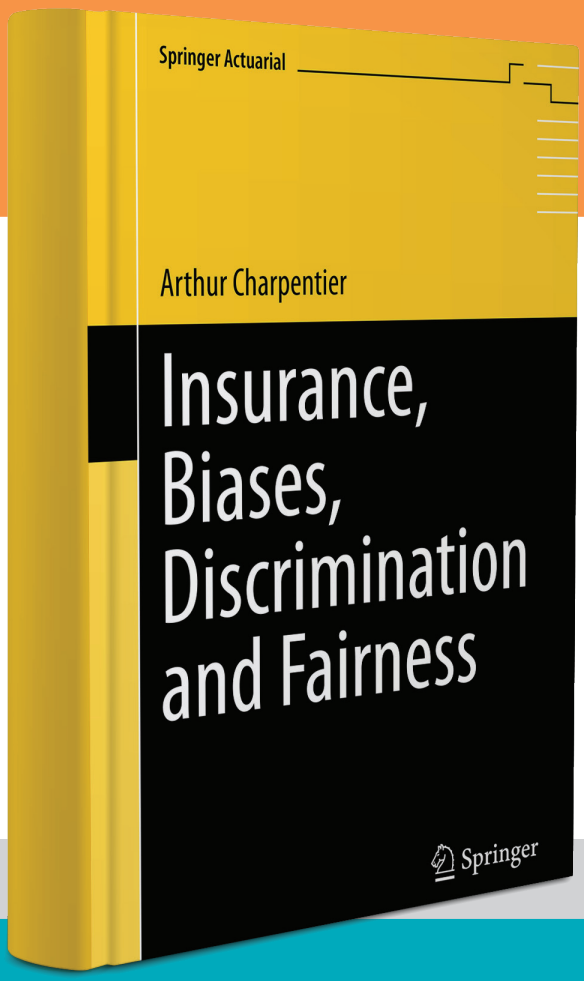

فراخوان ترجمه کتاب

پژوهشکده بیمه، به منظور کمک به گسترش دانش بیمه‌ای، ترجمه کتاب

Insurance, Biases, Discrimination and Fairness

را در دستور کار خود قرار داده است. لذا از کلیه اساتید، پژوهشگران، صاحبنظران و کارشناسان دعوت می‌شود که در صورت تمایل به ترجمه کتاب مذکور، کاربرگ درخواست ترجمه پیوست را به همراه سوابق علمی و اجرایی خود و ترجمه صفحات ذکر شده با ذکر عنوان کتاب، حداکثر تا تاریخ ۱۴۰۳/۰۷/۳۰ به آدرس ایمیل nashr@irc.ac.ir ارسال فرمایند.

معیارهای ارزیابی	امتیازات	ضریب
کیفیت ترجمه	میانگین امتیاز ۲ داور (حداکثر ۱۰)	۱
سوابق علمی	سوابق علمی مرتبط با موضوع کتاب: دکترا ۱۰ – ارشد ۸ – کارشناسی ٦ سوابق علمی غیرمرتبط: دکترا ٤ – ارشد ۳ – کارشناسی۲	۰.۲
سوابق تالیف/ترجمه کتاب	سوابق مرتبط با موضوع کتاب: حداکثر ۱۰ امتیاز براساس نرمال‌سازی سوابق غیرمرتبط : ۲۰ درصد امتیاز فوق	۰.٤
سابقه فعالیت تخصصی در حوزه بیمه	حداکثر ۱۰ امتیاز براساس نرمال‌سازی	۰.٤

کاربرگ درخواست ترجمه کتاب

عنوان کتاب:	Insurance, Biases, Discrimination and Fairness
ناشر: Springer	**سال نشر:** ۲۰۲۴

الف- اطلاعات عمومی

نام و نام خانوادگی	
شغل و سمت فعلی	
مرتبه علمی (ویژه اعضای هیات‌علمی)	
آخرین مدرک تحصیلی و رشته	
آدرس	
شماره تماس ثابت	
شماره تماس همراه	
پست الکترونیک	

ب- سابقه تألیف/ترجمه (حداقل ۳ عنوان از آثار خود را اعلام بفرمائید)

ردیف	عنوان کتاب/ترجمه	سال انتشار	ناشر

ج- سابقه اجرایی

ردیف	محل خدمت	مدت زمان خدمت

1.1.4 From Discrimination to Fairness

Humans have an innate sense of fairness and justice, with studies showing that even 3-year-old children have demonstrated the ability to consider merit when sharing rewards, as shown by Kanngiesser and Warneken (2012), as well as chimpanzees and primates (Brosnan 2006), and many other animal species. And given that this trait is largely innate, it is difficult to define what is "fair," although many scientists have attempted to define notions of "fair" sharing, as Brams et al. (1996) recalls. On the one hand "fair" refers to legality (and to human justice, translated into a set of laws and regulations), and in a second sense, "fair" refers to an ethical or moral concept (and to an idea of natural justice). The second reading of the word "fairness" is the most important here. According to one dictionary, fairness "*consists in attributing to each person what is due to him by reference to the principles of natural justice*." And being "just" raises questions related to ethics and morality (we do not differentiate here between ethics and morality).

This has to be related to a concept introduced in Feinberg (1970), called "*desert theory*," corresponding to the moral obligation that good actions must lead to better results. A student should deserve a good grade by virtue of having written a good paper, the victim of an industrial accident should deserve substantial compensation owing to the negligence of his or her employer. For Leibniz or Kant, a person is supposed to deserve happiness in virtue for being morally good. In Feinberg (1970)'s approach, "deserts" are often seen as positive, but they are also sometimes negative, like fines, dishonor, sanctions, condemnations, etc. (see Feldman (1995), Arneson (2007) or Haas (2013)). The concept of "desert" generally consists of a relationship among three elements: an agent, a deserved treatment or good, and the basis on which the agent is deserving.

We evoke in this book the "*ethics of models*," or, as coined by Mittelstadt et al. (2016) or Tsamados et al. (2021), the "*ethics of algorithms*." A nuance exists with respect to the "*ethics of artificial intelligence*," which deals with our behavior or choices (as human beings) in relation to autonomous cars, for example, and which will attempt to answer questions such as "*should a technology be adopted if it is more efficient?*" The ethics of algorithms questions the choices made "by the machine" (even if they often reflect choices—or objectives—imposed by the person who programmed the algorithm), or by humans, when choices can be guided by some algorithm.

Programming an algorithm in an ethical way must be done according to a certain number of standards. Two types of norms are generally considered by philosophers. The first is related to conventions, i.e., the rules of the game (chess or Go), or the rules of the road (for autonomous cars). The second is made up of moral norms, which must be respected by everyone, and are aimed at the general interest. These norms must be universal, and therefore not favor any individual, or any group of individuals. This universality is fundamental for Singer (2011), who asks not to judge a situation with his or her own perspective, or that of a group to which one belongs, but to take a "neutral" and "fair" point of view.

Proof See Feller (1957) or Ross (1972). □

This formula can be written simply in the case where two sets, two subgroups, are considered, for example, related to the gender of the individual,

$$\mathbb{E}(Y) = \mathbb{E}(Y|\text{woman}) \cdot \mathbb{P}(\text{woman}) + \mathbb{E}(Y|\text{man}) \cdot \mathbb{P}(\text{man}).$$

If Y denotes the life expectancy at the birth of an individual, the literal translation of the previous expression is that the life expectancy at birth of a randomly selected individual (on the left) is a weighted average of the life expectancies at birth of females and males, the weights being the respective proportions of males and females in the population. And as $\mathbb{E}(Y)$ is an average of the two,

$$\min\{\mathbb{E}(Y|\text{woman}), \mathbb{E}(Y|\text{man})\} \leq \mathbb{E}(Y) \leq \max\{\mathbb{E}(Y|\text{woman}), \mathbb{E}(Y|\text{man})\};$$

in other words, treating the population as homogeneous, when it is not, means that one group is subsidized by the other, which is called "actuarial unfairness," as discussed by Landes (2015), Frezal and Barry (2019), or Heras et al. (2020). The greater the difference between the two conditional expectations, the greater the unfairness. This "unfairness" is also called "cross-financing" as one group will subsidize the other one.

Definition 2.7 (Pure Premium (Heterogeneous Risks)) Let Y be the non-negative random variable corresponding to the total annual loss associated with a given policy, with covariates $\boldsymbol{x}$, then the pure premium is $\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$.

We use notation $\mu(\boldsymbol{x})$, also named "regression function" (see Definition 3.1). We also use notations $\mathbb{E}_Y[Y]$ (for $\mathbb{E}[Y]$) and $\mathbb{E}_{Y|\boldsymbol{X}}[Y|\boldsymbol{X} = \boldsymbol{x}]$ (for $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$) to emphasize the measure used to compute the expected value (and to avoid confusion). For example, we can write

$$\mathbb{E}_Y[Y] = \int_{\mathbb{R}} y f_Y(Y)\mathrm{d}y \text{ and } \mathbb{E}_{Y|\boldsymbol{X}}[Y|\boldsymbol{X} = \boldsymbol{x}] = \int_{\mathbb{R}} y f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})\mathrm{d}y$$
$$= \int_{\mathbb{R}} y \frac{f_{Y,\boldsymbol{X}}(y, \boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})}\mathrm{d}y.$$

The law of total expectations (Proposition 2.2) can be written, with that notation

$$\mathbb{E}_Y[Y] = \mathbb{E}_{\boldsymbol{X}}\big[\mathbb{E}_{Y|\boldsymbol{X}}[Y|\boldsymbol{X}]\big].$$

An alternative is to write, with synthetic notations $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y|\boldsymbol{X}]\big]$, where the same notation—$\mathbb{E}$—is used indifferently to describe the same operator on different probability measures.

For the pruning procedure, create a very large and deep tree, and then, cut some branches. Formally, given a large tree $\mathcal{T}_0$ identify a subtree $\mathcal{T} \subset \mathcal{T}_0$ that minimizes

$$\sum_{m=1}^{|\mathcal{T}|} \sum_{i:\mathbf{X}_i \in R_m} \left(Y_i - \widehat{Y}_{R_m}\right)^2 + \alpha|\mathcal{T}|,$$

where α is some complexity parameter, and $|\mathcal{T}|$ is the number of leaves in the subtree $\mathcal{T}$. Observe that it is similar to penalized methods described previously, used to get a tradeoff between bias and variance, between accuracy and parsimony.

Those classification and regression trees are easy to compute, to interpret. Unfortunately, those trees are rather unstable (even if the prediction is much more robust). The idea introduced by Breiman (1996a) consists in growing multiple trees to get a collection of trees, or a "forest," to improve classification by combining classifiers obtained from randomly generated training sets (using a "bootstrap" procedure—corresponding to resampling, with replacement), and then to aggregate. This will correspond to "bagging", which is an ensemble approach.

3.3.6 *Ensemble Approaches*

We have seen so far how to estimate various models, and then, we use some metrics to select "the best one." But rather than choosing the best among different models, it could be more efficient to combine them. Among those "ensemble methods," there will bagging, random forests, or boosting (see Sollich and Krogh (1995), Opitz and Maclin (1999) or Zhou (2012)). Those techniques can be related to "*Bayesian model averaging*", which linearly combines submodels of the same family, with the posterior probabilities of each model, as coined in Raftery et al. (1997) or Wasserman (2000), and "stacking", which involves training a model to combine the predictions of several other learning algorithms, as described in Wolpert (1992) or Breiman (1996c).

It should be stressed here that a "weak learner" is defined as a classifier that is correlates only slightly with the true classification (it can label examples slightly better than random guessing, so to speak). In contrast, a "strong learner" is a classifier that is arbitrarily well correlated with the true classification. Long story short, we discuss in this section the idea that combining "weak learner" could yield better results than seeking a "strong learner".

A first approach is the one described in Fig. 3.17: consider a collection of predictions, $\{\widehat{\boldsymbol{y}}^{(1)}, \cdots, \widehat{\boldsymbol{y}}^{(k)}\}$, obtained using k models (possibly from different families, GLM, trees, neural networks, etc.), consider a linear combination of those models, and solve a problem like

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \left\{ \sum_{i=1}^{n} \ell\left(y_i, \boldsymbol{\alpha}^\top \widehat{\boldsymbol{y}}_i\right) \right\}.$$

The receiver operating characteristic (ROC) curve is the curve obtained by representing the true-positive rates according to the false-positive rates, by changing the threshold. This can be related to the "discriminant curve" in the context of credit scores, in Gourieroux and Jasiak (2007).

Definition 4.19 (ROC Curve) The ROC curve is the parametric curve

$$\Big\{\mathbb{P}[m(\boldsymbol{X}) > t|Y = 0], \mathbb{P}[m(\boldsymbol{X}) > t|Y = 1]\Big\}, \text{ for } t \in [0, 1],$$

when the score $m(\boldsymbol{X})$ and Y evolve in the same direction (a high score indicates a high risk).

$$C(t) = \text{TPR} \circ \text{FPR}^{-1}(t),$$

where

$$\begin{cases} \text{FRP}(t) = \mathbb{P}[m(\boldsymbol{X}) > t|Y = 0] = \mathbb{P}[m_0(\boldsymbol{X}) > t] \\ \text{TPR}(t) = \mathbb{P}[m(\boldsymbol{X}) > t|Y = 1] = \mathbb{P}[m_1(\boldsymbol{X}) > t]. \end{cases}$$

In other words, the ROC curve is obtained from the two survival functions of $m(\boldsymbol{X})$ FPR and TPR (respectively conditional on $Y = 0$ and $Y = 1$). The AUC, the area under the curve, is then written as follows,

Definition 4.20 (AUC, Area Under the ROC Curve) The area under the curve is defined as the area below the ROC curve,

$$\text{AUC} = \int_0^1 C(t)\mathrm{d}t = \int_0^1 \text{TPR} \circ \text{FPR}^{-1}(t)\mathrm{d}t.$$

In Fig. 4.26, we can visualize on the left-hand side a classification tree, when we try to predict the gender of a driver using telematic information (from the `telematic` dataset), and on the right-hand side, ROC curves associated with three models, a smooth logistic regression (GAM), `adaboost` (boosting, GBM) and a random forest, trained on 824 observations, and ROC curves are based on the 353 observations of the validation dataset.

The AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. Indeed, assume for simplicity that the score (actually m_0 and m_1) has a derivative, so that the true-positive rate and the false-positive rate are given by

$$\text{TPR}(t) = \int_t^1 m_1'(x)\,\mathrm{d}x \text{ and } \text{FPR}(t) = \int_t^1 m_0'(x)\,\mathrm{d}x,$$

Box 6.6 (continued)
a person's age from their first name, indicates that "*the perceived age of a first name corresponds only weakly with the true average year of birth of people with that name*." Interpretation must be undertaken with caution when the study focuses on other characteristics of first names. Parents in different locations in the social circles choose different first names, Besnard and Grange (1993). The frequency of first names therefore varies with the social environment. Today, individuals called Anouk, Adèle, or Joséphine taken as a group have parents with more education than those called Anissa, Mégane, or Deborah. But this relationship depends on time, as some of the first names spread—following fashion—from one environment to another. Finally, when the characteristics associated with first names are linked to fluid, contextual identities, or assigned administratively from the outside, it is the addition of other variables that gives meaning to first names. Many studies (Tzioumis 2018 or Mazieres and Roth 2018) seek to exploit the information on ethnic, national or geographical origin contained in names and surnames, but they need, to start the investigation, a link between the first name and the variable studied. Directories from different countries, for example. The generalization of correlations between origin and first name to other countries or other populations must be made with caution.

Similarly, in "*are Emily and Greg more employable than Lakisha and Jamal* ?," Bertrand and Mullainathan (2004) randomly assigned African American or white-sounding names in resumes to manipulate the perception of race. "*White names*" received 50% more callbacks for interviews. Voicu (2018) presents the Bayesian Improved First Name Surname Geocoding (BIFSG) model to use first names to improve the classification of race and ethnicity in a mortgage-lending context, drawing on Coldman et al. (1988) and Fiscella and Fremont (2006). Analyzing data from the German Socio-Economic Panel, Tuppat and Gerhards (2021) show that immigrants with first names considered uncommon in the host country disproportionately complain of discrimination. When names are used as markers indicating ethnicity, it has been observed that highly educated immigrants tend to report perceiving discrimination in the host country more frequently than less educated immigrants. This phenomenon is referred to as the "discrimination paradox." Rubinstein and Brenner (2014) show that the Israeli labor market discriminates on the basis of perceived ethnicity (between Sephardic and Ashkenazi-sounding surnames). Carpusor and Loges (2006) analyzes the impact of first and last names on the rental market, whereas Sweeney (2013) analyze their impact on online advertising.

Definition 8.20 (Sufficiency (Barocas et al. 2017)) A model $m : \mathcal{Z} \to \mathcal{Y}$ satisfies the sufficiency property if $Y \perp\!\!\!\perp S \mid m(\boldsymbol{Z})$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$.

As discussed in Sect. 4.2 (and Definition 4.23) this property is closely related to calibration of the model. For Hedden (2021), it is the only interesting criterion to define fairness with solid philosophical grounds, and Baumann and Loi (2023) relates this criteria to the concept of "actuarial fairness" discussed earlier.

Definition 8.21 (Calibration Parity, Accuracy Parity (Kleinberg et al. 2016; Zafar et al. 2019)) Calibration parity is met if

$$\mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \texttt{A}] = \mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \texttt{B}], \ \forall t \in [0, 1].$$

We can go further by asking for a little more, not only for parity but also for a good calibration.

Definition 8.22 (Good Calibration (Kleinberg et al. 2017; Verma and Rubin 2018)) Fairness of good calibration is met if

$$\mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \texttt{A}] = \mathbb{P}[Y = 1 | m(\boldsymbol{X}) = t, S = \texttt{B}] = t, \ \forall t \in [0, 1].$$

In the context of a classifier, instead of conditioning on $m(\boldsymbol{X})$, we can simply use $\widehat{Y}$, as suggested in Chouldechova (2017).

Definition 8.23 (Predictive Parity (1)—Outcome Test (Chouldechova 2017)) We have predictive parity if

$$\mathbb{P}[Y = 1 | \widehat{Y} = 1, S = \texttt{A}] = \mathbb{P}[Y = 1 | \widehat{Y} = 1, S = \texttt{B}].$$

Note that if $\widehat{y}$ is not a perfect classifier ($\mathbb{P}[\widehat{Y} \neq Y] > 0$), and if the two groups are not balanced ($\mathbb{P}[S = \texttt{A}] \neq \mathbb{P}[S = \texttt{B}]$), then it is impossible to have predictive parity and equal opportunity at the same time. Note that positive predictive value is

$$\text{PPV}_s = \frac{\text{TPR} \cdot \mathbb{P}[S = s]}{\text{TPR} \cdot \mathbb{P}[S = s] + \text{FPR} \cdot (1 - \mathbb{P}[S = s])}, \ \forall s \in \{\texttt{A}, \texttt{B}\},$$

such that $\text{PPV}_{\texttt{A}} = \text{PPV}_{\texttt{B}}$ implies that either TPR or FPR is zero, and as negative predictive value can be written

$$\text{NPV}_s = \frac{(1 - \text{FPR}) \cdot (1 - \mathbb{P}[S = s])}{(1 - \text{TPR}) \cdot \mathbb{P}[S = s] + (1 - \text{FPR}) \cdot (1 - \mathbb{P}[S = s])}, \ \forall s \in \{\texttt{A}, \texttt{B}\},$$

such that $\text{NPV}_{\texttt{A}} \neq \text{NPV}_{\texttt{B}}$, and predictive parity cannot be achieved.

Continuing the formalism of Chouldechova (2017), Barocas et al. (2019) proposed an extension to predictive parity, with a distinction.

Chapter 10
Pre-processing

Abstract "Pre-processing" is about distorting the training sample to ensure that the model we obtain is "fair," with respect to some criteria (defined in the previous chapters). The two standard techniques are either to modify the original dataset (and to distort features to make them "fair," or independent of the sensitive attribute), or to use weights (as used in surveys to correct for biases). If there are poor theoretical guarantees, there are also legal issues with those techniques.

As we have seen previously, given a dataset $\mathcal{D}_n$, which is a collection of observations $(\boldsymbol{x}_i, s_i, y_i)$, it is possible to estimate a model $\widehat{m}$ (as discussed in Part II) and to quantify the fairness of the model (as discussed in Part III) based on appropriate metrics. Therefore, there are different ways of mitigating a possible discrimination, if any. The first one is to modify $\mathcal{D}_n$ (namely "pre-processing," described in this chapter), to modify the training algorithm ("in-processing," possibly in adding a fairness constraint to the objective function, as described in Chap. 8), or by distorting the trained model $\widehat{m}$ ("post-processing," as described in Chap. 9).

Pre-processing techniques can be divided into two categories. The first one modifies the original data, as suggested in Calmon et al. (2017) and Feldman et al. (2015), but it does not have many statistical guarantees (see Propositions 7.6 and 7.8, where we have seen that if we were able to ensure that $\boldsymbol{X}^\perp \perp\!\!\!\perp S$, we can still have $\widehat{Y} = m(\boldsymbol{X}^\perp) \not\perp\!\!\!\perp S$ – even $\widehat{Y} = m(\boldsymbol{X}^\perp) \not\perp S$)). Barocas and Selbst (2016) and Krasanakis et al. (2018) also question the legality of that approach where training data are, somehow, falsified. And another approach is based on reweighing, where instead of having observations with equal weights in the training sample, we adjust weights in the training sample (in the training function to be more specific, so it could actually be seen as an "in-processing" approach). Kamiran and Calders (2012) suggest simply having two weights, depending on whether s_i is either `A` or `B`, as well as Jiang and Nachum (2020). Heuristically, the idea is to amplify the error from an "underrepresented group" in the training sample, so that the optimization procedure can equally update a model for a different group.

A. Charpentier, *Insurance, Biases, Discrimination and Fairness*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-49783-4_10