L'équité de l'apprentissage machine en assurance

Laurence Barry, Arthur Charpentier

Working paper

26

L'équité de l'apprentissage machine en assurance

Laurence Barry et Arthur Charpentier

Février 2022

PARI

PROGRAMME DE RECHERCHE
SUR L'APPRÉHENSION DES RISQUES
ET DES INCERTITUDES

L'équité de l'apprentissage machine en assurance

Laurence Barry[1] et Arthur Charpentier[2, 3]

Technology is neither good nor bad; nor is it neutral[4]

Résumé

Depuis le début de leur histoire, les assureurs sont réputés utiliser des données pour classer et tarifer les risques. A ce titre, ils ont été assez tôt confrontés aux problèmes d'équité et de discrimination associées aux données. Pourtant, si cette question est récurrente, elle connait un regain d'importance avec l'accès à des données de plus en plus granulaires, massives et comportementales. Nous verrons ici comment les biais de l'apprentissage machine en assurance renouvellent ou transforment ce questionnement pour rendre compte des technologies et des préoccupations sociétales actuelles : paradoxalement, alors que la plupart de ces biais ne sont pas nouveaux, la recherche d'une équité pour les contrer, elle, se transforme.

Introduction

Dans leur analyse de l'impact des techniques actuarielles au cours des deux derniers siècles, Knights and Vurdubakis (1993) soutiennent que l'assurance *crée* le risque ou départage ce qui, dans l'incertitude, sera couvert par des mécanismes collectifs de ce qui restera du domaine de l'incertain. Le risque est la part quantifiée, modélisée, de l'incertitude ; c'est aussi sa part prise en charge par les institutions, assureurs ou Etat providence. Ainsi « the quantitative principles adopted by insurance (…) derive their particular rationality from the institution of *socially and historically specific modes of cognition and intervention* » (Knights et Vurdubakis 1993, 735, nos italiques).

Depuis le début de leur histoire, les assureurs quantifient le réel, fabriquant et utilisant des données pour classer et tarifer les risques. Cette pratique n'est pas anodine : l'assurance joue dans les sociétés industrialisées un rôle prépondérant

[1] Chaire PARI (ENSAE/Sciences Po)

[2] Université du Québec à Montreal (UQAM)

[3] Nous tenons à remercier Pierre François pour ses précieux commentaires.

[4] Kranzberg 1986

dans l'ouverture ou la fermeture d'opportunités de vie (Horan 2021; Baker et Simon 2002). A ce titre, les assureurs ont été assez tôt confrontés aux questions d'équité associées aux données. Mettre en évidence un biais, c'est adopter une position critique par rapport à un calcul, et éclairer la dimension politique que cache sa prétendue objectivité. Mais la contestation est, elle aussi, culturellement et historiquement située : on verra dans la mise en perspective historique des biais en assurance, proposée ici en première partie, comment ils sont reformulés, précisés, réorganisés avec les évolutions culturelles et technologiques.

Il convient d'emblée de lever une ambiguïté : dans le langage courant, la discrimination renvoie toujours à une injustice ; la pratique actuarielle, elle, repose sur une discrimination d'ordre statistique, qui se veut neutre et objective. Pour Charpentier (2021), l'usage du terme en statistique remonte probablement aux premiers travaux de Ronald Fisher dans les années 20, dont le but était de différencier et classer (en anglais, *discriminate*) deux espèces de fleurs suivant les caractéristiques de leurs pétales. Il n'empêche qu'en assurance, pour les raisons évoquées plus haut, toute classification comme discrimination statistique est susceptible d'être perçue comme une discrimination sociale.

Dès 1909, le régulateur du Kansas pose les contours d'une pratique éthique de la classification, dans le but de protéger les petits souscripteurs d'une assurance incendie qui payaient des primes beaucoup plus élevées que les gros industriels, pour un risque identique. Il définit ainsi une tarification comme « non inéquitablement discriminatoire » (*not unfairly discriminatory*), si elle traite de la même manière des personnes semblables (Frezal et Barry 2020; M. J. Miller 2009). C'est sur la base de ce principe que l'usage de certains paramètres en tarification assurantielle ont été contestés dans le courant du 20e siècle, aboutissant à une typologie assez précise des biais liés à un traitement classique des données.

L'émergence des données massives et des nouveaux algorithmes bouscule conceptuellement la pratique assurantielle : là où la mutualisation était la règle, on entend de plus en plus qu'une « individualisation » du risque serait plus éthique, et rendue possible par ces nouvelles technologies. On verra cependant dans la deuxième partie que la discussion des biais induits et des diverses notions d'équité algorithmique remet au goût du jour et renouvelle, avec des points de rupture et de continuité, des débats plus anciens liés à la discrimination en assurance. La dernière partie discute enfin des enjeux éthiques des données massives et de l'apprentissage machine, ou *machine learning*, pour l'assurance.

Assurance, biais et équité : une approche historique

La pratique de tarification avant l'apprentissage machine

L'assurance consiste en la mise en commun de l'incertitude : la contribution de chacun permet la compensation des accidents survenus aux plus malchanceux. Dans sa forme la plus grossière, la prime de risque assurantielle est l'espérance mathématique des dommages de l'accident, calculée sur le groupe en question. Si la concurrence ne joue pas entre assureurs, on peut très bien se contenter d'un tarif unique, moyenne du risque pour l'ensemble de la population. Mais la concurrence

fait craindre l'antisélection : en réduisant la prime des meilleurs risques, l'assureur A peut les attirer à lui, bonifiant ainsi son portefeuille aux dépens de ses concurrents qui eux feront des pertes. La segmentation, qui consiste à distinguer des groupes porteurs de risques voisins, devient donc très vite la règle du jeu.

Cette segmentation a consisté pendant très longtemps en la création de classes supposées homogènes, sur lesquelles le risque est estimé en moyenne (Charpentier, Denuit, et Elie 2015). Le travail de l'actuaire était donc, avant tout calcul, celui du choix des variables, choix qui dictait une homogénéité projetée sur le monde, et ce de deux façons : dans le choix de ce qui est ignoré d'une part, puisque ce qui n'est pas collecté contient des différences qui ne seront pas vues ; dans la catégorisation de ce qui est collecté d'autre part, qui conduit là encore à écraser des différences potentielles.

Petit à petit, émerge l'idée d'une « tarification parfaite », dans laquelle la classe tarifaire ne comporterait que des risques parfaitement identiques. En reprenant la formalisation de Denuit and Charpentier (2004), et si l'on admet que thêta est la variable qui caractériserait parfaitement le risque :

	Assuré	Assureur
Perte	$\mathbb{E}[Y\|\Theta]$	$Y - \mathbb{E}[Y\|\Theta]$
Perte moyenne	$\mathbb{E}[Y]$	0
Variance	$\mathrm{Var}\left[\mathbb{E}[Y\|\Theta]\right]$	$\mathrm{Var}\left[Y - \mathbb{E}[Y\|\Theta]\right]$

La variance sur le portefeuille est ainsi distribuée entre les assurés qui paient des primes proportionnelles à leur risque (capté par thêta) et l'assureur qui porte la variance résiduelle, inexpliquée par thêta. Dans les années 80, et avec une notation similaire, De Wit et Van Eeghen (1984) estiment que les capacités croissantes de collecte de données et de calcul des ordinateurs, permettent d'envisager l'affinement de la part expliquée de la variance (et les primes segmentées), diminuant ainsi celle portée par l'assureur.

Cependant, le paramètre thêta, supposé caractériser le risque de façon parfaite, n'est en réalité jamais connu. Cette incertitude est le fondement même de l'assurance : si on cherche, par exemple, à modéliser et valoriser des garanties en cas de décès, on peut estimer de manière plus fine la probabilité de décès (certains ayant 1 chance sur 10,000 et d'autres 1 chance sur 1,000 de mourir), mais il demeure impossible de prédire *qui* va décéder dans l'année. Cette incertitude résiduelle fondamentale reste irréductiblement à la charge de l'assureur, créant ce que De Wit et Van Eeghen (1984) appelle une solidarité purement probabiliste (couverte par la loi des grands nombres). La classification est alors repensée comme un moyen d'approcher thêta : on ne cherche plus seulement à contrer l'antisélection avec des classes de plus en plus fines, mais on interprète ce travail comme une approximation de théta par les paramètres de tarification, comme un moyen de faire converger la variance non expliquée par le modèle vers la variance minimale du portefeuille. Ainsi la pratique actuarielle ne change pas, même si son sens évolue.

C'est dans ce cadre d'ajustement qu'apparait la notion de biais : dans l'hypothèse où un calcul exact du risque est possible, il se produit si la classification est imparfaite et conduit à mal tarifer certains groupes, créant des transferts croisés entre assurés (Walters 1981; De Pril et Dhaene 1996).

La critique d'une tarification biaisée

A partir des années 60 aux Etats-Unis, la classification des risques est remise en cause, sur deux aspects spécifiques. Dans le contexte de lutte pour les droits civiques des noirs tout d'abord, c'est la pratique du « red lining » ou d'exclusion de certaines zones géographiques des portefeuilles assurés qui est montrée du doigt. Puis, à la fin des années 70, les mouvements féministes tentent de contrer l'usage du sexe dans la tarification (Horan 2021). Ainsi dans l'affaire Manhart en 1978 puis dans l'affaire Norris en 1983, la Cour suprême a jugé que l'utilisation du paramètre homme/femme dans la tarification d'un régime de retraite à prestations définies était illégale, car cela violait les principes d'égale opportunité et d'avancement individuel (Avraham 2017; Austin 1983; Horan 2021).

Les débats autour des actions de groupe menées alors mettent en lumière les différents aspects de ce que l'on peut appeler une « tarification biaisée ». Nous proposons dans cette section une description théorique de ces débats, débouchant sur une typologie des biais assurantiels pré-machine learning. En réalité, le sens de la critique dépend fortement de l'hypothèse que l'on fait sur le monde et la nature du risque. Or cette dernière reste le plus souvent implicite dans les arguments avancés, ce qui rend les débats parfois assez opaques.

La classification peut en effet être pensée tout d'abord comme une méthode de répartition ex-ante des coûts futurs, toujours plus ou moins arbitraire. Dans cette hypothèse, le travail de quantification (du statisticien ou de l'actuaire) est critiqué car il se nourrirait d'une vision du monde toujours subjective, dictée par un contexte historique et culturel spécifique (Desrosières 2008). Pour l'exemple, une lecture des manuels de formation à la souscription des années 70 révèle une description des femmes comme peu fiables, instables dans leur travail, incapables de prendre des décisions financières de façon autonome et dépendantes de leur partenaire masculin pour vivre (Horan 2021, 174), justifiant l'usage du paramètre homme/femme dans la tarification.

Glenn (2000) fait alors remarquer que, comme le dieu romain Janus, le processus de sélection des risques d'un assureur a en réalité deux visages : Il y a d'un côté le visage des chiffres, des tables actuarielles, et des statistiques, qui se posent comme objectives et rationnelles. Mais de l'autre il y a le visage des récits, du caractère et du jugement subjectif. Pour Glenn, l'actuaire crée un mythe dans lequel les décisions apparaissent comme objectives alors qu'elles reposent sur beaucoup de subjectivité, de préjugés et de stéréotypes. Ces derniers sont visibles en amont des tables actuarielles, dans les histoires que se racontent les techniciens de l'assurance (actuaires et souscripteurs), et qui les amènent à privilégier telle variable plutôt que telle autre. En effet, comme la collecte des données se fait encore sous forme de questionnaires, elle est à la fois coûteuse et nécessairement contrainte en volume. Elle est aussi contingente à ce que l'on peut techniquement et/ou historiquement

mesurer, ce qui induit une certaine instabilité dans la classification. Comme le dit Baker (2002): « While some 'low risk' individuals may believe that they are benefited by risk classification, any particular individual is only one technological innovation away from losing his or her privileged status » (voir aussi Frezal and Barry 2020).

De plus, dans les controverses autour de la classification, Horan (2021, 170–71) montre que les paramètres de tarification évoluent aussi pour répondre aux contraintes réglementaires, politiques ou sociales : « 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 ». L'histoire des biais en assurance est de fait aussi l'histoire de ce qui est perçu comme acceptable ou inacceptable dans une société donnée. Dans cette perspective, des classifications alternatives peuvent être également efficaces, mettant en évidence la marge de manœuvre, mais aussi l'arbitraire des décisions laissées aux praticiens, eux-mêmes guidés ou contraints par le contexte dans lequel ils évoluent :

> 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).

Dans l'affaire Manhart, l'un des juges met ainsi en avant la fluidité culturelle et historique de ce qui est perçu comme légitime, à la fois comme classification ex-ante mais aussi comme explication ex-post du modèle :

> Habit, rather than analysis, makes it seem acceptable and natural to distinguish between male and female, alien and citizen, legitimate and illegitimate; for too much of our history there was the same inertia in distinguishing between black and white. But *that sort of stereotyped reaction may have no rational relationship--other than pure prejudicial discrimination-to the stated purpose for which the classification is being made* (cité dans Simon 1988, 796, nos italiques).

Pour Schauer (2003) il conviendrait de distinguer deux types de stéréotypes. Certaines généralisations sont totalement infondées : des généralisations sur la base du signe astrologique de la personne, par exemple, relèvent de purs préjugés. Mais d'autres ont un fondement statistique, lorsque la probabilité d'avoir un caractère y sachant x est significativement différente du cas où l'on ne sait rien. Dans cette perspective, l'usage du paramètre homme/femme reste légitime car statistiquement fondé pour estimer une probabilité de décès ou d'accident automobile. Serait alors légitime toute classification sur la base de variables effectivement corrélées au risque que l'on cherche à modéliser.

Works (1977) met cependant en garde contre les « variables de procuration », par opposition aux « vraies variables » du risque. Ces dernières seraient plus difficiles à obtenir, du coup remplacées par de simples approximations - d'où la porte ouverte à tous les biais dans la tarification et la souscription :

> Although the core concern of the underwriter is the human characteristics of the risk, *cheap screening indicators are adopted as surrogates for solid information* about the attitudes and values of the prospective insured (...) The invitations to underwriters *to introduce prejudgments and biases and to indulge amateur psychological stereotypes are apparent*. Even generalized underwriting texts include occupational, ethnic, racial, geographic, and cultural characterizations certain to give offense if publicly stated (Works 1977, 471, nos italiques).

Dans les actions de groupe menées aux Etats-Unis contre l'usage du paramètre homme/femme, c'est cette approche qui est adoptée par les plaignantes. Leur argument principale est en effet que la corrélation observée entre coût des sinistres en assurance automobile et sexe du conducteur est due au moindre kilométrage parcouru par les femmes ; c'est le kilométrage qui est la variable causale, donc légitime, et non le sexe qui n'est qu'une approximation biaisée de cette dernière (Horan 2021). On voit bien pourtant que l'hypothèse sous-jacente est ici radicalement différente : il existerait de « vraies variables » du risque, qui expliqueraient les accidents de façon causale, toutes les autres étant invalides : ce n'est pas l'éthique de la classification qui est remise en cause mais sa mauvaise application.

Le problème de ce type d'argument vient de ce qu'il est très difficile d'établir l'existence d'une causalité directe, et que par conséquent il s'agit plus d'un jugement, là encore subjectif, qu'une vraie preuve scientifique : on pourrait alors dire qu'une causalité est encore de l'ordre du récit, lorsque celui-ci est accepté comme scientifiquement et/ou politiquement valide. Pour l'exemple, lorsque Hoffman à la fin du 19e siècle met en évidence une corrélation entre la durée de vie et la couleur de la peau, il en déduit l'existence d'une causalité innée liée à la race noire et qui la rend plus risquée, là où d'autres auraient cherché les causes environnementales et sociales expliquant la plus grande mortalité des noirs (Heen 2009, 377).

Dans le même ordre d'idée, Simon (1988, 795-6) soutient que causalité ou corrélation finalement importent peu lorsqu'il s'agit de lutter contre une discrimination sociale flagrante : sur cette base, l'usage du paramètre discriminant contribue à naturaliser la différence de traitement (social) et donc à ancrer dans la réalité la discrimination – d'où le besoin de l'éliminer des variables autorisées dans le traitement statistique. Nous reviendrons plus loin sur les problèmes posés par le décalage qui se crée alors entre la réalité sociale et les statistiques collectées.

Les critiques de la classification classique déterminent ainsi une typologie des différents biais possibles, que l'on retrouvera de façon modifiée dans les méthodes d'apprentissage machine :

- Les biais de type 1 sont liés à des classes qui ne reflèteraient pas la réalité du risque, mais seraient motivés par de purs préjugés (critique qui ne remet pas en question le principe du bien-fondé de la classification). Justifiée en amont par le mythe d'une causalité des signes astrologiques sur les accidents par exemple, une classification zodiacale se révèlerait à l'usage comme « biaisée », au sens trivial où le modèle est faux ;
- Les biais de type 2 sont liés à des classes qui reflètent une réalité statistique avérée (une corrélation avec le risque, donc un modèle exact) mais non

causale, ce qui rend leur usage suspect d'un parti-pris et d'un choix arbitraire. C'est le cas par exemple du paramètre homme/femme. Là aussi on admet le bien-fondé d'une classification qui s'appuierait uniquement sur des variables causales, mais la corrélation seule ne donne pas lieu à une explication acceptable ;

- Les biais de types 3 sont liés à des classes qui reflètent une réalité statistique et causale, mais qui est elle-même le fait de discriminations sociales en amont. Dans ce cas, le modèle est exact mais la classification est intrinsèquement nuisible car elle reproduit et ancre dans la réalité une situation contre laquelle il faut lutter.

Il est intéressant de noter ici que cette typologie ne décrit pas intrinsèquement telle ou telle variable, mais la façon dont on se représente les bais. On verra dans la section qui suit comment certains paramètres peuvent être reconnus comme des variables causales et acceptables socialement à un moment donné de leur histoire, pour basculer ensuite dans la catégorie des variables corrélées à une cause plus profonde et/ou dans celle des variables protégées.

Les variables protégées

Pour répondre au troisième type de biais et prévenir ou remédier à la discrimination sociale, on peut choisir d'interdire l'usage de certaines variables, dites protégées ou sensibles. Cette section décrit de manière non exhaustive les controverses associées à l'usage historique de quelques paramètres controversés, créant des biais d'un type ou d'un autre ; nous verrons notamment que la sensibilité d'une variable est contingente au contexte culturel. En Europe, les données protégées concernent aujourd'hui notamment les croyances religieuses, le sexe, l'orientation sexuelle, l'engagement syndical, l'appartenance ethnique, la situation médicale, les condamnations et infractions pénales, les données biométriques, les informations génétiques.

Même si la réglementation est moins stricte dans d'autres pays, des projets voient le jour ailleurs pour limiter l'usage de certains facteurs. Aux États-Unis par exemple, le projet de loi sur l'interdiction de la discrimination en matière d'assurance automobile, ou PAID Act (*Prohibit Auto Insurance Discrimination Act*), présenté en juillet 2019 au Congrès, interdit à un assureur automobile de tenir compte de certains facteurs non directement liés à la conduite lorsqu'il détermine la prime d'assurance, ou son admissibilité (Metz 2020; Watson Coleman 2019). Le genre du conducteur, sa situation d'emploi, son code postal, son secteur de recensement, son état civil, et son score de crédit deviendraient des variables interdites.

L'origine ethnique

Alors qu'en France la collecte et l'usage statistique de l'origine ethnique des individus reste un sujet polémique, ils sont assez répandus aux États-Unis. En assurance-vie, Bouk (2015) décrit ainsi comment, à la fin du 19e siècle, les assureurs faisaient payer la même prime à tout le monde mais réglaient les sinistres de façon différenciée suivant la couleur de la peau (voir aussi Heen 2009):

> Industrial insurers operated a high-volume business; so to simplify sales they charged the same nickel to everyone. The home office then calculated benefits according to actuarially defensible discriminations, by age initially and then by race. In November 1881, Metropolitan decided to mimic Prudential, allowing policies to be sold to African Americans once again, but with the understanding that black policyholders' survivors only received two-thirds of the standard benefit (Bouk 2015, 34).

Plusieurs États adoptent alors des lois anti-discrimination. Ainsi au cours de l'été 1884, l'État du Massachusetts promulgue une loi interdisant de faire « 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 » (cité par Wiggins 2013, 68). La loi exigeait également que les assureurs paient des indemnités complètes aux assurés afro-américains (Wiggins 2013). Pour contrer la loi, Frederick L. Hoffman, soutenu par Prudential Life Insurance, publie en 1896 un ouvrage démontrant, sur la base de statistiques (partielles et donc biaisées), la mortalité plus élevée des Noirs américains (Bouk 2015, 49-52; Heen 2009, 377). Les assurer au même tarif que les Blancs serait statistiquement inéquitable, soutenait-il ; ne pas les assurer était donc la seule manière de se conformer à la loi, qui rendait de fait les Noirs américains non-assurables.

Le sujet reste d'actualité pendant la majeure partie du 20e siècle,[5] même si la couleur de peau disparait des tables actuarielles après la seconde guerre mondiale. Pour Heen (2009, 364), c'est moins la législation - qui interdisait l'usage de l'origine raciale depuis la fin du 19e siècle- que les leçons de la guerre et du nazisme qui conduisent les assureurs à bannir le paramètre : « Change came from a form of collective action by life insurance industry professional groups, which was achieved only after a fundamental rethinking of race, a 'change in the habit of the public mind' that led to reconsideration of long-established commercial practice » (Heen 2009, 399).

Ce ban est cependant purement cosmétique puisqu'il est remplacé par la zone géographique, non pas comme variable causale mais comme « proxy », ou variable de procuration (Works 1977). En effet, l'origine ethnique peut être inférée avec une assez grande précision du lieu d'habitation des assurés potentiels. Une enquête commissionnée par l'Etat fédéral dans les années 60 met ainsi en évidence la pratique systématique de « red-lining » (Austin 1983; Horan 2021) : de nombreuses institutions financières, dont des compagnies d'assurance, refusent de desservir des

[5] Heen (2009) soutient qu'il est possible que dans certains Etats du Sud des Etats-Unis, d'anciennes polices d'assurance vie issues de la période Jim Crow (i.e. faisant usage de la race comme paramètre de tarification) soient encore en vigueur aujourd'hui.

zones géographiques à prédominance afro-américaine, conduisant à une détérioration des services et des infrastructures dans certaines villes :

> Without insurance, banks and other financial institutions will not, and cannot make loans. New houses cannot be built. Existing houses cannot be repaired. New businesses cannot be started. Existing ones cannot expand, or even survive. Thus, without insurance an area deteriorates. Its services, goods, and jobs, the lifeline of the city, diminish. Communities without insurance are communities without hope (Cité dans Horan 2021, 140–41).

Tout se passe ici comme si une variable corrélée mais devenue inacceptable, est remplacée par une autre de même type : on choisit volontairement un biais de type 2 pour contourner la législation qui visait à éviter un biais de type 3. Plus récemment, une étude en assurance automobile a révélé que les quartiers à prédominance afro-américaine paient en moyenne 70% de plus que les autres quartiers (Heller 2015). Larson et al. (2017) ont effectué une analyse similaire par code postal pour les principales compagnies d'assurance à travers les États-Unis, confirmant l'existence d'un écart, quoique plus faible : la surprime est en moyenne de 10% en responsabilité civile automobile pour les codes postaux associés à des populations minoritaires. En réponse, l'association étatsunienne des assureurs *Property Casualty (Property Casualty Insurers Association of America)* affirmait que « insurance rates are color-blind and solely based on risk » (cité dans Larson et al. 2017).

Discrimination Homme/Femme : aléa subi ou volontaire

Comme évoqué plus haut, c'est sur l'usage du paramètre homme/femme que les premières controverses autour de la classification actuarielle se sont faites jour. En Europe, une directive de 2004 visait à réduire les écarts entre les sexes dans l'accès à tous les biens et services, mais une dérogation permettait aux assureurs de fixer des prix fondés sur le paramètre homme/femme, à condition qu'ils fournissent des données actuarielles et statistiques permettant d'établir qu'il constitue un facteur objectif d'évaluation du risque. En 2011, soit trente ans après les controverses étatsuniennes, la Cour de justice des Communautés européennes a annulé cette exception, rendant l'usage du paramètre homme/femme caduque pour toutes les classifications (Schmeiser, Störmer, et Wagner 2014; Rebert et Van Hoyweghen 2015), au motif qu'il ne serait que corrélé avec la cause réelle de l'accident (donc biais de type 2).

Dans sa décision, la juge fait par ailleurs la distinction entre deux types de variables, pointant ce qui pourrait être considérée comme une classification équitable, une fois éliminées les variables non significatives et les variables non causales : « A l'instar de la race et de l'origine ethnique, le sexe est lui aussi une caractéristique inséparable de la personne de l'assuré sur laquelle *celui-ci n'a pas la moindre influence* » (CURIA 2010, nos italiques). Cette distinction renvoie à ce que Dworkin (1981) appelle « brute and option luck » : les aléas que l'on dira volontaires sont liés à des choix personnels (*option luck*) et peuvent être imputés à l'individu ; les aléas subis, causés par des éléments sur lesquels l'individu n'a aucune prise (*brute luck*), devant

eux être pris en charge par la collectivité (et donc protégés et éliminés de la tarification).[6]

Discrimination par l'âge

A première vue, l'âge comme le sexe ou l'appartenance éthnique est une donnée personnelle sur laquelle l'individu n'a pas prise et devrait donc être, au regard du texte précédent, proscrit des tables actuarielles. Il y a cependant une différence majeure qui en fait un paramètre acceptable. En effet, toujours dans les conclusions de la juge on trouve :

> S'il est vrai que l'âge est, lui aussi, une caractéristique indissociablement liée à la personne, *tout homme traverse différentes tranches d'âge au cours de son existence*. C'est ainsi que, si les primes et prestations d'assurance sont calculées différemment en fonction de l'âge, cela ne permet pas de craindre, en soi, que l'assuré s'en trouve lésé *en tant que personne*. Quiconque peut, au cours de sa vie, bénéficier, en fonction de son âge, de produits d'assurance plus ou moins avantageux pour lui (CURIA 2010, nos italiques).

L'âge n'est pas une caractéristique discrète et immuable. Comme le dit Macnicol (2006), l'âge « n'est pas un club dans lequel on naît » ; nous nous attendons à passer par les différentes étapes de la vie et la vieillesse est « un club » que nous rejoindrons très probablement un jour. Par conséquent, dans une perspective temporelle longue, le traitement différentiel en fonction de l'âge ne génère pas nécessairement des inégalités entre les personnes : « une société qui discrimine sans relâche les gens en raison de leur âge peut encore les traiter de manière égale tout au long de leur vie (…) Le tour de chacun <d'être discriminé> viendra » (Gosseries 2014). Cet argument n'est pourtant valide que si la discrimination reste de même nature au cours du temps ; il admet l'hypothèse, le plus souvent vraie, que la société discrimine toujours les personnes plus âgées. Mais les normes sociales et les mécanismes de solidarité évoluent. Ainsi la retraite, financée par répartition en France, fonctionne grâce une solidarité intergénérationnelle qui fait que le poids des retraites pèse sur les actifs. Or cet équilibre dépend de la pyramide des âges, qui est dynamique et fait qu'au cours du temps, certaines générations se trouvent pénalisées par rapport à d'autres. Après-guerre, lors de la mise en place du régime, du fait de l'espérance de vie et de l'âge légal de départ à la retraite, beaucoup de cotisants ne bénéficièrent jamais de leur retraite, par exemple. Plus tard, l'allongement de la durée de cotisation et la baisse du niveau des retraites montrent bien que cette notion de compensation au cours de la vie ne fonctionne pas toujours.

Par ailleurs, en suivant la distinction (dans le biais de type 2) entre variable causale et variable simplement corrélée, il n'est pas évident que l'âge soit la cause de la mortalité. L'âge permet d'inférer assez précisément l'état de santé de la personne, cause réelle du décès, mais variable protégée. L'usage de l'âge pourrait donc introduire des biais dans les modèles. Analysant ainsi un arrêt de la cour d'appel de 2008, Mercat-Bruns (2020) conclut que « le législateur a pris soin d'opérer une distinction entre l'âge et l'état de santé. Il ne peut dès lors être procédé à un

[6] Cette distinction n'est pas toujours simple à établir : voir Charpentier, Barry, et James (2020) pour une discussion dans le cas des catastrophes naturelles.

amalgame entre ces deux motifs en considérant que l'âge avancé induit nécessairement une santé défaillante ».

Discrimination des fumeurs

La responsabilité du tabagisme dans la genèse des cancers (en particulier du poumon) a été longue à établir. Le rôle cancérigène du tabac a été suspecté au lendemain de la Première Guerre mondiale, et le lien entre certains cancers et le tabagisme est établi par les assureurs dès 1930 (Patterson 1989). Hoffman – le statisticien de Prutential responsable des tables de mortalité raciales - collecte notamment des statistiques à partir de 1915 et conclut: « smoking habits unquestionably increase the liability to cancer of the mouth, the throat, the oesophagus, the larynx and the lungs » (Hoffman 1931, 67).

Les premières quantifications interviennent après-guerre, avec notamment les travaux de Johnston (1945) qui présentent des tables de mortalités comparant non-fumeurs et fumeurs (« moderate » et « heavy »). Des études de grande envergure ont lieu dans les années 1950 et 1960 : Doll et Hill (1964) confirment ainsi le lien entre tabagisme et cancer. Dans un contexte purement actuariel, il faut attendre les années 80, toujours aux Etats-Unis, pour que les tables de mortalité homologuées tiennent compte de cette variable. Une « task force » est ainsi créée par la Société des Actuaires en 1982 pour proposer une correction aux tables de mortalité en vigueur, grâce à un facteur fumeur/non-fumeur (Society of Actuaries 1982; G. H. Miller et Gerstein 1983). Dans les années 80, des travaux similaires seront menés en Europe (Benjamin et Michaelson 1988). En France, le paramètre est rarement utilisé jusqu'à aujourd'hui, même si l'impact sur la mortalité est avéré.

Le facteur a en réalité longtemps fait polémique : ainsi Fisher (1958) met en garde contre l'amalgame entre corrélation et causation. Pour lui, les études montrent toutes l'existence d'une corrélation avec le cancer du poumon, mais ne prouvent pas que le tabagisme en est la cause : « it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others » (Fisher 1958). Il s'évertue à montrer qu'en réalité l'inclination à fumer est génétique et que c'est aussi cette configuration génétique qui est à l'origine du surplus de cancers dans la population des fumeurs. Dans la perspective de ce papier, le débat autour du tabagisme proposé par Fisher met en avant deux types de biais potentiel : le fait que le facteur fumeur/non-fumeur ne serait pas un facteur causal (biais de type 2) ; le fait que si la cause est génétique, alors elle tombe dans la catégorie des variables sur lesquelles l'individu n'a pas prise et devrait donc être bannie des tarifs pour des raisons d'équité (indépendamment du fait que le traitement des données génétiques est interdit au titre du RGPD).

Les scores de crédit

En Amérique du Nord, diverses entreprises telles qu'Experian, Equifax et TransUnion, tiennent des registres des activités d'emprunt et de remboursement d'une personne. La société FICO (Fair Isaac Corporation) a mis au point une formule (non connue) calculant, sur la base de ces registres, un score, fonction de la dette et du crédit disponible (Guseva et Rona-Tas 2001). Ce score est utilisé non

seulement par les établissements de crédit, mais aussi à l'embauche (Bartik et Nelson 2019). C'est aussi le cas en assurance, même si M. J. Miller et Smith (2003) insistent sur le fait que ce n'est pas le score FICO proprement dit qui est utilisé, mais un autre indicateur dérivé des registres d'activité, lui-même prédicteur d'accidents.

Ces usages font cependant aujourd'hui débat car ils créent un cercle vicieux d'appauvrissement des plus pauvres (O'Neil 2016). François (2021) met également en avant l'aspect auto-réalisateur de la pratique puisqu'un mauvais score augmente le coût du crédit et par conséquent les chances de ne pouvoir le rembourser. En assurance, le régulateur américain s'est récemment penché sur l'équité de la pratique. Dans la perspective de ce papier, il a notamment cherché à mettre en évidence l'existence d'un lien causal entre ce score basé sur les données de crédit et l'occurrence d'un accident ; il est alors apparu que ce que cet indicateur permettait de prédire était le dépôt d'une demande d'indemnisation, et non les accidents, c'est-à-dire qu'il fonctionnait comme un indicateur socio-économique de l'assuré et non pas comme un indicateur de risque (Kiviat 2019).

Les enjeux de l'apprentissage machine pour les biais en assurance

Les techniques de segmentation décrites dans la partie précédente, mises en place dans le courant du 20e siècle, impliquaient toujours l'intervention lourde de l'actuaire ou du statisticien, de ce fait responsable des biais de ses modèles. A partir des années 2000, avec l'émergence des données massives, on a de plus en plus recours à des techniques d'apprentissage machine, qui permettraient de remplacer l'humain par la machine dans un certain nombre de tâches : peut-on en déduire pour autant que les biais seront réduits ? Rien n'est moins sûr. Et qu'en est-il, plus précisément en assurance ?

L'apprentissage machine : qu'est-ce qui change ?

Mesurer l'impact des données massives en assurance est peut-être plus difficile que dans d'autres domaines. D'un côté, comme toute autre organisation, les assureurs sont amenés à modifier leurs pratiques pour intégrer les nouvelles sources de données devenues accessibles, les capacités de calcul accrues et les nouveaux algorithmes. De l'autre pourtant, ces techniques apparaissent souvent comme la continuation d'une pratique de segmentation presque séculaire (Swedloff 2014). De plus, certaines études montrent qu'à ce jour les modèles de tarification n'ont pas profondément changé, ni que de nouveaux produits n'ont émergé, suite par exemple à l'apparition des boitiers télématiques (Barry et Charpentier 2020; François et Voldoire 2022). L'étude ci-dessous tient donc plus d'une analyse de ce que les nouveaux modèles rendent possibles, même si le basculement, en assurance, n'a pas (encore ?) été observé en pratique.

La première modification qui vient nourrir l'apprentissage machine est l'apparition des données massives. A la différence de l'ère précédente, ces données ne sont plus obtenues via des questionnaires qui impliquaient un travail en amont dans le choix de ce que l'on voulait collecter et suivant quelle codification (Desrosières 2008). Aujourd'hui ces données sont obtenues via des senseurs, des objets connectés, ou

sont nativement numériques car procédant d'actions en ligne – autant de sources qui ne demandent pas a priori d'intervention humaine. A la grande différence des données issues de questionnaires, ces données sont par ailleurs le plus souvent des données comportementales : pour les senseurs, et en se limitant à l'assurance des particuliers, on peut citer les boitiers télématiques qui collectent en continu la position, la vitesse et l'accélération du véhicule (Barry et Charpentier 2020), ou les bracelets connectés mesurant des données biométriques de leurs porteurs (Lupton 2014; 2016).

La deuxième modification majeure tient aux capacités de calculs des ordinateurs, sans commune mesure avec la génération précédente. Ainsi lorsque De Wit et Van Eeghen (1984) évoquent la possibilité d'affiner la segmentation, ils s'appuient sur l'idée que « *with the help of computer*s it has become possible to make thorough risk analyses, and consequently to arrive at further premium differentiation » (De Wit et Van Eeghen 1984, 155, nos italiques): c'est l'existence même des ordinateurs qui, dans les années 80, changent la donne par rapport à une époque antérieure où les calculs étaient pratiquement manuels (Barry 2020). Aujourd'hui, ce sont les capacités de calcul qui permettent le traitement de bases de données beaucoup plus importantes.

Enfin, l'apprentissage machine quant à lui, permet d'automatiser une partie des tâches, notamment celle du choix des variables significatives, ce qui démultiplie le nombre de variables dont on peut tenir compte. Les modèles deviennent ainsi plus complexes, sans nécessairement changer de nature. C'est le cas par exemple avec les modèles de « price optimization », qui permettent de tenir compte dans la tarification non seulement du risque de l'assuré, mais aussi de sa sensitivité au prix de l'assurance et de sa propension à résilier son contrat. Ces modèles posent des problèmes nouveaux en terme d'équité, puisque ce serait les clients les plus loyaux qui se trouveraient pénalisés au profit d'assurés dans la même classe de risque, mais eux plus sensibles au prix de leur assurance (Frees et Huang 2021).

Un saut conceptuel a lieu en revanche avec les algorithmes d'apprentissage profond (ou *deep learning*, pris ici comme une catégorie d'apprentissage machine). LeCun, Bengio, et Hinton (2015) caractérisent en effet l'apprentissage profond par sa capacité à inférer seul les relations potentielles entre variables, antérieurement imposées aux données par l'analyste : « the key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure ».

Mis en perspective avec la partie précédente, l'apprentissage machine semble donc a priori lever les biais de type 1 et 2 qui résultaient des préjugés et stéréotypes de l'actuaire dans son choix et sa codification des variables. L'accès récent aux données comportementales semblent par ailleurs répondre au besoin de distinguer entre variables décrivant un choix conscient de l'assuré (son comportement) et celles relevant de caractéristiques intrinsèques auxquelles il ne peut rien.

Il n'est donc pas anodin que dans les conclusions de l'affaire Test-Achats, la juge invoque l'imprécision des statistiques lorsque le risque serait en réalité individuel :

> L'espérance de vie des assurés (…) est fortement influencée par des éléments économiques et sociaux ainsi que *par les habitudes de vie* de tout un chacun

> comme, par exemple, la nature et l'intensité de *l'activité professionnelle*, l'environnement familial et social, les *habitudes alimentaires*, la consommation de denrées d'agrément (…) ou de drogues, les *activités de loisirs*, la *pratique du sport* (CURIA 2010).

Autrement dit, le comportement individuel plus que les statistiques, nécessairement agrégées, permet de déterminer le risque (donc individuel). La juge semble encourager la tarification sur la base des données de comportements rendues accessibles par les senseurs :

> The gender directive gave telematics insurance products an added push. Since insurers are no longer able to differentiate by gender in their rating – and previously gender had quite an influence on the premium charged – telematics offers the opportunity to see how someone really drives. Hence, you get products now that are designed to reward better drivers, regardless of their gender. The likelihood is that, certainly at the younger end, more of the better drivers will be female (Munich Re, cité dans Meyers 2018).

Les conclusions de l'affaire Test-Achats vont ainsi dans le sens du mouvement général qui conçoit le risque comme associé au mode de vie, donc comme individuel, et non plus comme déterminé sur la base de classes statistiques (Rebert et Van Hoyweghen 2015). C'est aussi l'objet du projet de loi étatsunien PAID, qui stipule que tout paramètre *non directement lié à la conduite* devrait être interdit dans la tarification du risque automobile (Metz 2020). On retrouve ici, transposée à l'assurance, l'utopie que les algorithmes actuels seraient capables de personnaliser les décisions au niveau individuel, là où leurs ancêtres se contentaient de travailler sur des moyennes sur des sous-groupes (Lury et Day 2019; Moor et Lury 2018).

Les biais de l'apprentissage machine en assurance

Paradoxalement pourtant, la bascule de classes statistiques aux données massives (et comportementales) dans un but de personnalisation et d'ajustement du risque ne fait qu'exacerber les biais évoqués en première partie, en modifiant leur nature.

Refléter la réalité des risques

Les compagnies d'assurances s'appuient de plus en plus sur des données de sources externes : le plus fréquemment, ce sont des actions en ligne, que ce soient des factures, des transactions, des courriers électroniques, des photos, des flux de clics, des journaux, des requêtes de recherche, des dossiers médicaux, etc (Charpentier 2021). Dans l'utopie un peu mythologique du big data, ces données seraient enfin devenues exhaustives, permettant de rendre compte de la réalité de façon plus riche : sans compromis lié à l'échantillonnage, sans contraintes de volume de données (Mayer-Schönberger et Cukier 2014), et sans la réduction du réel due au travail de quantification (Desrosières 2008).

Mais l'un des problèmes essentiel lié à ces données tient au fait qu'elles résultent de l'observation et non d'expériences construites ad hoc (Rosenbaum 2017; Charpentier 2021) – d'où un biais d'échantillon. La méthode des expériences randomisées, formalisée par Ronald Fisher en 1935 (Fisher 1971), consiste à constituer deux groupes grâce auxquels on mesure l'effet d'une variable (un traitement en médecine, par exemple). Ces deux groupes doivent avoir des caractéristiques proches (à défaut d'être identiques), la proximité se mesurant par une série de co-variables (l'âge, le genre, etc). L'un des groupes, dit de contrôle,

n'est pas soumis au traitement, ce qui permet, *toute chose égale par ailleurs*, d'observer l'effet du traitement sur le deuxième groupe. Il y a deux points supplémentaires à respecter dans cette approche : les patients ne doivent pas savoir quel traitement ils ont eu ; le traitement doit être choisi de manière totalement indépendante des patients. Ce sont des points qui opposent fondamentalement les données résultant d'expériences randomisées aux données observationnelles (Leigh 2018; Rosenbaum 2017).

Or dans la plupart des cas, les bases de données massives sont observationnelles puisqu'elles résultent du comportement des gens indépendamment de toute expérience ad hoc. C'était déjà le point mis en avant par Ronald Fisher dans la polémique autour du tabagisme : pour lui, sans expérience randomisée qui permettrait de comparer des populations identiques, l'observation d'une corrélation entre tabagisme et cancer ne prouve rien. Comme sa thèse est qu'un facteur génétique explique à la fois tabagisme et cancer, il constitue des groupes de jumeaux homozygotes (des populations donc identiques) et met en évidence les comportements similaires des jumeaux sur les deux facteurs (Fisher 1958).

La réalité présentée par les données massives est elle-aussi filtrée, même si cela n'est plus le fait du statisticien qui construit sa base (boyd et Crawford 2012). Pour Barocas et Selbst (2016), les populations en marge de l'économie formelle et des activités en ligne sont nécessairement sous-représentées dans ces données, créant des risques de discrimination. De plus, comme l'écrivait déjà Desrosières (1993) à propos des statistiques classiques, « les indicateurs quantitatifs rétroagissent sur les acteurs quantifiés » . Le biais de rétroaction intervient lorsque les acteurs intègrent le fait qu'un paramètre fait l'objet d'une mesure pour la tarification : ils modifient alors leur comportement afin d'agir en retour sur le paramètre mesuré. Ce biais est magnifié lorsqu'il s'agit de variables comportementales. L'actuaire, qui n'a pas lui-même construit les bases de données auxquelles il a à présent accès, conçoit parfois mal ces limites.

Un autre biais d'échantillon est lié à l'auto-sélection (Charpentier 2021). Cette situation se retrouve de plus en plus, notamment dans les fichiers administratifs. En effet, jusque très récemment, les données étaient stockées automatiquement mais, paradoxalement, le RGPD - dont le but essentiel est la protection des données personnelles-, conduit à un biais lié au non-consentement de certains : il est en effet aujourd'hui possible à ceux qui en font la demande de voir leurs données supprimées. Ce concept de *opting-out* peut fortement biaiser les données conservées.

Un autre exemple devenu classique en data-science est celui évoqué par Caruana et al. (2015), qui porte plus sur la non-exhaustivité des données : des chercheurs dans les années 90 avaient entrainé un réseau de neurones profond pour classer les patients entre faible et haut risque de pneumonie, afin de limiter les admissions à l'hôpital. Le modèle était extrêmement précis sur les données d'entraînement, mais les résultats étaient contre-intuitifs sur les patients asthmatiques, classés comme risque faible. Un examen plus détaillé montrait que les patients asthmatiques étaient en pratique traités beaucoup plus rapidement, justement en raison de leur risque de mortalité très élevé en cas de pneumonie. Leur risque faible résultait donc d'un

traitement différencié (les deux populations n'étaient pas identiques) dont l'algorithme ne pouvait tenir compte.

Corrélation vs. causation : l'efficacité opaque des nouveaux algorithmes

L'opacité des nouveaux algorithmes, souvent décriée, est mise en balance dans les débats avec leur précision accrue en comparaison des modèles classiques (Breiman 2001). En 2017, lors de l'un des premiers débats à la conférence NeurIPS,[7] il avait été souligné que « 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 ». Lors de la conférence, Yann LeCun souligne que lorsqu'on leur présentait deux modèles (l'un parfaitement interprétable et précis à 90%, l'autre une boîte noire ayant une précision supérieure de 99%), les gens choisissaient toujours le modèle plus précis. Il en conclut que « people don't really care about interpretability but just want some sort of reassurance from the working model ». L'interprétabilité n'est pas importante si l'on est convaincu que le modèle fonctionne bien dans les conditions dans lesquelles il est censé fonctionner.

Pour Napoletani, Panza, et Struppa (2011, 3), il s'agit d'un nouveau paradigme scientifique, ouvrant la voie vers une science devenue « agnostique ». À ce titre, dans un article désormais célèbre Anderson (2008) parle de la « fin des théories » pour caractériser la nouvelle approche qui conduit à renoncer à mettre en évidence des liens de causalité entre les variables (donc à expliquer le phénomène) de l'analyse :

> Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise. But *faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete* (Anderson 2008, nos italiques).

Dans la perspective de ce papier, cela voudrait dire renoncer à l'explicitation des relations entre variables, causales ou corrélées : la machine produit un score, suffisamment précis pour justifier l'abandon d'une interprétation par les données en entrée. Quelques exemples, encore assez rares, de cette approche boite noire existent en assurance. On peut penser à des applications en gestion de la fraude (quand envoyer un expert ?) ou en marketing (qui solliciter, ou quel produit proposer ?) : dans ces domaines, l'interprétation est largement négligée au profit de la rentabilité. Pour la tarification, certains algorithmes de reconnaissance d'images peuvent inférer des facteurs de risque. Par exemple, Kita-Wojciechowska et Kidziński (2019) proposent de prédire la fréquence d'accident automobile à partir d'images-satellite du lieu d'habitation du conducteur ; ou encore Shikhare (2021) calcule un score de santé sur la base d'une photo-portrait de la personne.

---

[7] Appelé «*The Great AI Debate : Interpetability is necessary for machine learning* », opposant Rich Caruana et Patrice Simard (pour) à Kilian Weinberger et Yann LeCun (contre), https://youtu.be/93Xv8vJ2acI.

Corriger les discriminations sociales : le paradoxe des variables protégées dans un environnement de données massives

Dans les modèles boite noire, les biais dus aux préjugés et stéréotypes du statisticien que Works (1977) essayait d'éviter en recommandant l'usage de variables causales, seraient écartés puisque c'est l'algorithme qui établit des liens (inconnus) entre les variables devenues pléthoriques. En réalité on sait aujourd'hui qu'au contraire les préjugés, stéréotypes et autres discriminations se retrouvent dans les données elles-mêmes, donc bien en amont du jugement des statisticiens : au-delà des biais d'échantillon évoqués plus haut, c'est vraiment la nature des données qui est en cause (Caliskan, Bryson, et Narayanan 2017).

De plus, alors que dans les modèles classiques on pouvait espérer corriger les biais en interdisant l'usage de certaines variables dites protégées, la colinéarité de ces variables avec d'autres, facialement neutres, dans les données massives rend cette « protection » illusoire : « 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, nos italiques).

Pour Prince et Schwarcz (2019) la discrimination par procuration, évoquée déjà pour les modèles classiques, est magnifiée par les nouveaux algorithmes. Alors qu'elle était intentionnelle par le passé (puisqu'une décision humaine présidait au choix des variables – par exemple le red-lining), la discrimination par procuration devient non-intentionnelle. Suivant la distinction établie par Barocas et Selbst (2016), la discrimination par procuration ne résulte plus d'un traitement consciemment différencié des segments protégés (*disparate treatment*), mais elle fait partie des discriminations transparentes dont on ne perçoit que les effets ex-post (*disparate impact*). Ce phénomène est inévitable, en particulier lorsqu'une variable directement liée au phénomène (une variable causale) est absente des données. C'est ce qui se passe dans l'exemple de Caruana : la variable causale est le temps de traitement depuis les premiers symptômes, qui explique la meilleure survie des asthmatiques. Mais cette variable ne fait pas partie de la base, du coup l'algorithme « apprend » que les asthmatiques sont moins risqués. C'est aussi ce qui menace de se produire lorsque l'on exclut des variables causales mais protégées car reflétant un aléa subi – par exemple les données génétiques en assurance santé (Prince et Schwarcz 2019, 1264) :

> 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 et Schwarcz 2019, 1274).

On risque alors de créer des algorithmes qui associent le visionnage de certains programmes télévisés à un facteur de risque en santé !

Pour lutter contre ce phénomène, Williams, Brooks, et Shmargad (2018) montrent que paradoxalement, il ne faut pas interdire la collecte et l'usage des variables protégées, mais au contraire s'en servir comme moyen de piloter la non-discrimination. C'est de cette manière par exemple que les Britanniques appréhendent les données ethniques, par opposition à la France :

> La collecte des données religieuses et ethniques ne pose plus problème depuis une trentaine d'années. Au contraire : ces données sont considérées par les Britanniques issus des minorités (on utilise pour les désigner l'acronyme BAME, pour « Black, Asian and minority ethnic ») comme un puissant outil d'action politique « positive » (Ducourtieux 2021)

Ce point illustre un problème supplémentaire lié à l'interdiction d'usage ou de collecte de certaines variables, dans le cas inverse à celui analysé par Williams, Brooks, et Shmargad (2018), c'est-à-dire celui où l'information véhiculée par la variable n'est pas captée par d'autres. On peut en effet alors se trouver devant un paradoxe de Simpson: la donnée manquante conduit à des conclusions erronées parce que l'information disponible n'est pas suffisamment granulaire ou captée par les autres variables (Bickel, Hammel, et O'Connell 1975; Charpentier 2021; Alipourfard, Fennell, et Lerman 2018).

Apprentissage machine et équité assurantielle : collective ou individuelle ?

L'équité est l'autre versant de la médaille qui détermine ce qui est perçu comme biais ou discrimination dans un contexte culturel et historique donné. Dans quelle mesure les technologies disponibles influent sur cette conception de la justice ? Est-ce que, notamment, l'environnement des données massives et des nouveaux algorithmes modifient la conception de l'équité assurantielle ?

Pour Thiery et Schoubroeck (2006), les juristes et les actuaires ont des conceptions fondamentalement différentes de l'équité. L'équité assurantielle et la segmentation reposeraient sur une vision collective de l'équité, alors que l'équité juridique met en avant les droits individuels. Juridiquement, le droit à l'égalité de traitement est octroyé à une personne en sa qualité d'individu, qui ne peut être traité différemment en raison de son appartenance à tel groupe ou tel groupe. Mais cette vision s'oppose fondamentalement à l'approche actuarielle qui, historiquement, analyse les risques et calcule les primes en termes collectifs (Ewald 2011).

Ainsi dans sa décision sur l'affaire Norris, le juge maintient qu'une classification statistiquement valide (qu'il s'agisse d'un lien causal ou d'une corrélation) n'en fait pas une classification légitime. En réalité, aucune classification ne peut l'être puisque « 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 » (cité dans Horan 2021, 187). Pour l'individu auquel on l'impose, la classification constitue toujours une discrimination, dite statistique (Binns 2018), ou une généralisation arbitraire de l'individu à un groupe.

Pour Simon (1988) et Horan (2021), l'adoption de ce point de vue individuel par le juge a contribué à renforcer l'effacement du principe de solidarité pourtant au cœur de la pratique assurantielle. L'assurance repose en effet sur la mise en commun de l'incertitude et s'appuie sur le voile d'ignorance qui met les uns et les autres à égalité devant l'aléa (Ewald 1986). Mais une fois posée l'existence d'un risque individuel qu'il faudrait approcher, par la classification puis par les nouveaux algorithmes, la tarification devient un exercice mathématique d'optimisation et de

minimisation de la variance portée par l'assureur. La « personnalisation » associée aux nouveaux algorithmes devient en assurance « l'individualisation » du risque (Barry et Charpentier 2020). Ainsi cette distinction entre équité assurantielle-collective et équité juridique-individuelle tend à disparaître (Barry 2020). Car même si les actuaires n'ont pas fondamentalement bouleversé leur pratique, la notion d'équité assurantielle semble évoluer avec les nouvelles technologies. Dans les produits télématiques, le risque n'est plus présenté comme une incertitude mise en commun, mais comme un choix individuel. Le comportement de chacun devrait déterminer sa prime, et non plus des données démographiques agrégées. L'équité dans ce cas consiste à ajuster la prime au comportement individuel, pour que chacun paie suivant « son » risque (Meyers et Van Hoyweghen 2018). Dans cette perspective, le biais statistique évoqué par le juge dans l'affaire Norris prend une importance renouvelée.

Mais cette individualisation, si elle a lieu, est problématique à plus d'un titre ; il n'est pas évident tout d'abord que le résultat soit équitable pour tous les assurés concernés. Elle conduirait en effet à des tarifications plus disparates, avec des primes pour les individus les plus risqués qui risqueraient de devenir inabordables, les excluant de fait de la communauté assurée (Charpentier, Barry, et Gallic 2020). Il n'est pas non plus évident d'autre part que le machine learning puisse résoudre cette tension entre équité individuelle et collective.

Face à l'opacité des modèles et aux biais sociaux embarqués dans les données massives, l'équité algorithmique émerge ainsi comme une nouvelle discipline (Kusner et Loftus 2020). On retrouve dans cette littérature la tension entre point de vue individuel ou collectif au cœur des questionnements actuels sur l'individualisation du risque en assurance. L'exactitude (mathématique) d'un algorithme se mesure en général à partir d'une matrice de confusion, qui permet d'observer les erreurs par type – faux négatifs et faux positifs. Mais la minimisation simultanée de ces erreurs n'est pas toujours possible, voire souhaitable, pour plusieurs raisons. En effet, faux positifs et faux négatifs ne sont pas comparables d'un point de vue éthique : la condamnation d'un innocent n'a pas la même « valeur » que la libération d'un coupable. Ainsi suivant le contexte, il faudra choisir de minimiser l'une ou l'autre forme d'erreur.

Les choses se compliquent encore lorsque l'on tient compte des variables protégées. Pessach et Shmueli (2020) distinguent alors entre des indicateurs d'équité collectifs ou individuels. Les indicateurs collectifs visent à assurer la parité entre groupes, protégé ou non. On peut ainsi tenter de s'assurer que les fréquences de prédiction (exacte ou positive) soient égales sur les deux groupes ; ou mesurer les taux de faux positifs et faux négatifs séparément sur les deux groupes et vérifier qu'ils sont voisins. Ce n'était pas le cas, par exemple, pour l'algorithme étatsunien COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*, ou algorithme prédictif d'aide à la décision des juges sur le risque de récidive). Il avait en effet un taux de faux positifs (faussement classés récidivistes) beaucoup plus élevés pour les noirs, et un taux de faux négatifs plus élevé pour les blancs, avec des précisions égales sur les deux groupes (Kleinberg, Mullainathan, et Raghavan 2016). Les indicateurs individuels, eux, visent à s'assurer que des individus similaires (hors variable protégée) obtiennent un score similaire. Kusner et Loftus (2020)

définissent de la sorte l'équité « contrefactuelle », qui consiste à comparer les scores de deux observations identiques sur lesquelles seule la variable protégée prend une valeur différente. Cette technique permet de répondre aussi précisément que possible à la question « que se serait-il passé si seul l'attribut protégé avait été différent ? », qui pour Pearl et Mackenzie (2018) constitue un test de causalité (ici du tarif). Tous les auteurs s'accordent sur le fait que ces différents indicateurs ne peuvent pas être optimisés simultanément, conduisant à de nécessaires compromis dépendants du contexte (Kleinberg, Mullainathan, et Raghavan 2016; Pessach et Shmueli 2020).

Le ban du paramètre homme/femme par la directive européenne exemplifie ces dilemmes en assurance : soit on ignore la variable, mais alors si une différence statistique existe elle réapparaîtra au travers d'autres variables, colinéaires au paramètre interdit et par conséquent la moyenne sur les hommes et les femmes restera différente ; soit au contraire on utilise cette variable pour maintenir des moyennes identiques mais alors toutes choses égales par ailleurs, le tarif variera en fonction du sexe de la personne. On ne pourra jamais maintenir la parité entre les groupes et assurer l'équité contrefactuelle. Pour Charpentier (2021, 148), interdire l'usage de la variable protégée est contre-productif car « dans la plupart des cas réalistes, non seulement la suppression de la variable sensible ne rend pas les modèles de régression équitables, mais au contraire, une telle stratégie est susceptible d'amplifier la discrimination ».

Conclusion

L'équité assurantielle est une notion dynamique, dont on a vu ici qu'elle dépendait de contextes historiques, culturels et techniques divers. En pleine ère industrielle, on s'appuyait sur le voile d'ignorance pour justifier des couvertures très larges en termes de solidarité, et sur l'idée de l'égalité du plus grand nombre face à une adversité mal connue. Cette équité était critiquée par les libéraux qui y voyaient une incitation à la licence. Dans le courant du 20e siècle, avec les capacités croissantes de collecte et de calcul se mettent en place des modèles segmentés, qui assoient l'assurance sur la classification des risques, perçus comme des groupes homogènes de personnes qui se ressemblent. A partir des années 80, des controverses se font jour autour de l'usage de telle ou telle variable, controverses qui constituent le lit des critiques actuelles concernant les biais et les discriminations associées à l'apprentissage machine.

L'examen de cette histoire permet d'identifier quelques familles principales de biais, dans les pratiques traditionnelles de classification puis leur déclinaison dans les algorithmes de machine-learning. On distingue ainsi avant tout les critiques qui admettent la classification dans son principe, mais conteste l'usage de telle ou telle variable. Ces critiques sont de deux ordres :

- On critique tout d'abord des variables reflétant les préjugés du statisticien qui choisit de les collecter pour créer son modèle, même lorsqu'elles n'ont aucun lien avec le phénomène à étudier. C'est le cas de la couleur de peau aux Etats-Unis dans les produits d'assurance-vie à la fin du 19e siècle. Ce type de biais disparait en principe avec les données

massives : étant nativement numériques, elles court-circuitent le travail de quantification de la période précédente. Mais on s'est rendu compte, au cours de ces vingt dernières années, que les préjugés ont la vie dure et que les discriminations sociales se retrouvent dans les données. Un usage aveugle de l'apprentissage machine conduirait alors à reproduire ces biais dans les modèles.

- Une autre forme de discrimination mise au jour dans les années 60-80 et que l'on retrouve magnifiée avec les nouveaux algorithmes tient à l'usage de variables corrélées sans être causales : ainsi l'usage des paramètres homme/femme, le score de crédit ou le critère fumeur/non-fumeur ont provoqué des controverses dont certaines se poursuivent jusqu'à aujourd'hui. La solution préconisée par les critiques, sûrement inopérable en pratique, serait de se limiter à des variables purement causales. Cette exigence de causalité avérée est totalement abandonnée dans les algorithmes de machine-learning, dont certains disent qu'ils signent l'avènement d'un nouvel *episteme* : ils se contentent en effet de mettre en évidence des corrélations entre les données en entrée, sans même expliciter ces liens. Ceci conduit à un biais nouveau lié à ces techniques, celui de leur opacité, même s'il est la contrepartie d'une plus grande précision.

Une autre grande famille de critiques rejette peu ou prou la classification :

- Dans les modèles classiques, l'équité de l'assurance exigeait que certaines variables causales soient exclues de l'analyse parce qu'elles reflètent un aléa subi et non choisi par la personne : l'usage de données génétiques en assurance santé par exemple est interdit dans la plupart des pays. Dans ce cas, l'assurance est perçue comme un moyen non plus de refléter le risque mais, en éliminant la variable des modèles, de le faire porter par l'ensemble de la population assurée. La solution de l'élimination des variables protégées, si elle est effective dans les modèles traditionnels, est beaucoup plus difficile à mettre en œuvre avec les données massives et l'apprentissage machine, respectivement parce que les variables protégées sont captées via leur colinéarité avec d'autres, et que l'opacité des algorithmes rend la mise en évidence de ces discriminations plus complexe.
- Plus fondamentalement, une critique légaliste opposait traditionnellement les droits de l'individu à la classification, soit encore une approche individuelle de l'équité à celle collective portée par l'assurance, mettant en avant le biais statistique induit par la réduction, nécessairement arbitraire, d'un individu aux données d'une classe. Avec les données massives dont certaines sont comportementales, l'utopie est de résoudre ce biais, en personnalisant et individualisant les modèles. Mais là aussi, la promesse n'est pas tenue : les théoriciens de l'équité algorithmique mettent en avant l'impossibilité d'optimiser les algorithmes sur divers critères simultanés, dont aucun ne peut être a priori préféré à un autre. Dans le contexte assurantiel, l'équité individuelle menace cependant de conduire à des tarifs de plus en plus différenciés, donc inabordables pour certaines personnes classées très risquées.

Faut-il alors en rester aux bonnes vieilles tables de tarification, pour lesquelles tous les paramètres sont explicites, connus à l'avance et par là-même ouverts à la contestation ? Un peu comme toute théorie scientifique se doit d'être falsifiable, une tarification se devrait d'être transparente afin d'être contestable.

Références

Alipourfard, Nazanin, Peter G. Fennell, et Kristina Lerman. 2018. « Can you Trust the Trend? Discovering Simpson's Paradoxes in Social Data ». In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 19-27. WSDM '18. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3159652.3159684.

Anderson, Chris. 2008. « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete ». *Wired*, 2008. https://www.wired.com/2008/06/pb-theory/.

Austin, Regina. 1983. « The Insurance Classification Controversy ». *University of Pennsylvania Law Review* 131 (3): 517. https://doi.org/10.2307/3311844.

Avraham, Ronen. 2017. « Discrimination and Insurance ». SSRN Scholarly Paper ID 3089946. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3089946.

Baker, Tom. 2002. « Containing the Promise of Insurance: Adverse Selection and Risk Classification ». SSRN Scholarly Paper ID 322581. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.322581.

Baker, Tom, et Jonathan Simon. 2002. « Embracing Risk ». In *Embracing Risk: The Changing Culture of Insurance and Responsibility*, 1-25. University of Chicago Press.

Barocas, Solon, et Andrew D. Selbst. 2016. « Big Data's Disparate Impact Essay ». *California Law Review* 104: 671-732.

Barry, Laurence. 2020. « Insurance, Big Data and Changing Conceptions of Fairness ». *European Journal of Sociology / Archives Européennes de Sociologie* 61 (2): 159-84. https://doi.org/10.1017/S0003975620000089.

Barry, Laurence, et Arthur Charpentier. 2020. « Personalization as a promise: Can Big Data change the practice of insurance? » *Big Data & Society*. https://journals.sagepub.com/doi/full/10.1177/2053951720935143.

Bartik, Alexander, et Scott Nelson. 2019. « Deleting a Signal: Evidence from Pre-Employment Credit Checks ». Working Paper 2019-137. Chicago: Becker Friedman Insitute - University of Chicago. https://www.ssrn.com/abstract=3498458.

Benjamin, B., et R. Michaelson. 1988. « Mortality Differences between Smokers and Non-Smokers ». *Journal of the Institute of Actuaries* 115 (3): 519-25. https://doi.org/10.1017/S0020268100042797.

Bickel, P. J., E. A. Hammel, et J. W. O'Connell. 1975. « Sex Bias in Graduate Admissions: Data from Berkeley ». *Science*, février. https://doi.org/10.1126/science.187.4175.398.

Binns, Reuben. 2018. « Fairness in Machine Learning: Lessons from Political Philosophy ». In *Conference on Fairness, Accountability and Transparency*, 149-59. PMLR. http://proceedings.mlr.press/v81/binns18a.html.

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

boyd, dana, et Kate Crawford. 2012. « Critical Questions for Big Data ». *Information, Communication and Society* 15 (5): 662-79. https://doi.org/10.1080/1369118X.2012.678878.

Breiman, Leo. 2001. « Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author) ». *Statistical Science* 16 (3): 199-231. https://doi.org/10.1214/ss/1009213726.

Caliskan, Aylin, Joanna J. Bryson, et Arvind Narayanan. 2017. « Semantics Derived Automatically from Language Corpora Contain Human-like Biases ». *Science*, avril. https://doi.org/10.1126/science.aal4230.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, et Noemie 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*, 1721-30. KDD '15. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2783258.2788613.

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

Charpentier, Arthur, Laurence Barry, et Ewen Gallic. 2020. « Quel avenir pour les probabilités prédictives en assurance ? » *Annales des Mines - Realites industrielles* 2020 (1): 74-77.

Charpentier, Arthur, Laurence Barry, et Molly James. 2020. « Insurance against Natural Catastrophes: Balancing Actuarial Fairness and Social Solidarity ». Working Paper 22. Paris: Chaire PARI.

Charpentier, Arthur, Michel M. Denuit, et Romuald Elie. 2015. « Segmentation et Mutualisation, les deux faces d'une même pièce ». *Risques*, nº 103: 19-23.

CURIA. 2010. « Test-Achats Conclusions de l'Avocat General ». Court of Justice of the European Union. https://curia.europa.eu/juris/document/document_print.jsf?docid=82589&text=&dir=&doclang=FR&part=1&occ=first&mode=lst&pageIndex=0&cid=7912963#Footnote36.

De Pril, Nelson, et Jan Dhaene. 1996. *Segmentering in verzekeringen*. Leuven: KUL. Departement toegepaste economische wetenschappen.

De Witt, G.W., et J. Van Eeghen. 1984. « Rate Making and Society's Sense of Fairness ». *ASTIN Bulletin*, nº 14:2: 151-64.

Denuit, Michel, et Arthur Charpentier. 2004. *Mathématiques de l'assurance non-vie: Principes fondamentaux de théorie du risque*. ECONOMICA edition. Paris: ECONOMICA.

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

Doll, Richard, et Austin Bradford Hill. 1964. « Mortality in Relation to Smoking: Ten Years' Observations of British Doctors ». *British Medical Journal* 1 (5396): 1460-67.

Ducourtieux, Cecile. 2021. « Les statistiques ethniques au Royaume-Uni, un outil essentiel pour lutter contre les inégalités ». *Le Monde.fr*, 22 avril 2021. https://www.lemonde.fr/economie/article/2021/04/22/les-statistiques-ethniques-au-royaume-uni-un-outil-essentiel-pour-lutter-contre-les-inegalites_6077646_3234.html?xtor&&M_BT=36351134033493#x3D;EPR-33281062-[la-lettre-eco]-20210422-.

Dworkin, Ronald. 1981. « What is Equality? Part 2: Equality of Resources ». *Philosophy & Public Affairs* 10 (4): 283-345.

Ewald, François. 1986. *L'Etat Providence*. Grasset.

———. 2011. « Omnes et Singulatim. After Risk ». *Carceral Notebooks* 7: 77-107.

Fisher, Ronald A. 1958. « Cancer and Smoking ». *Nature* 182 (4635): 596-596. https://doi.org/10.1038/182596a0.

———. 1971. *The Design of Experiments*. New York: Macmillan Pub Co.

François, Pierre. 2021. « Catégorisation, individualisation. Retour sur les scores de crédit ». Working Paper 24. Paris: Chaire PARI. https://www.chaire-pari.fr/wp-content/uploads/2021/10/WP-24-categorisation-individualisation.pdf.

François, Pierre, et Theo Voldoire. 2022. « The revolution that did not happen. Telematics and car insurance in the 2010s ». Working Paper 26. Paris: Chaire PARI.

Frees, Edward W. (Jed), et Fei Huang. 2021. « The Discriminating (Pricing) Actuary ». *North American Actuarial Journal* 0 (0): 1-23. https://doi.org/10.1080/10920277.2021.1951296.

Frezal, Sylvestre, et Laurence Barry. 2020. « Fairness in Uncertainty: Some Limits and Misinterpretations of Actuarial Fairness ». *Journal of Business Ethics* 167 (1): 127-36. https://doi.org/10.1007/s10551-019-04171-2.

Glenn, Brian J. 2000. « The Shifting Rhetoric of Insurance Denial ». *Law & Society Review* 34 (3): 779-808. https://doi.org/10.2307/3115143.

———. 2003. « Postmodernism: The Basis of Insurance ». *Risk Management & Insurance Review* 6 (2): 131-43. https://doi.org/10.1046/J.1098-1616.2003.028.x.

Gosseries, Axel. 2014. « What Makes Age Discrimination Special? A Philosophical Look at the ECJ Case Law ». *Netherlands Journal of Legal Philosophy* 43 (1): 59-80. https://doi.org/10.5553/NJLP/221307132014043001005.

Guseva, Alya, et Akos Rona-Tas. 2001. « Uncertainty, Risk, and Trust: Russian and American Credit Card Markets Compared ». *American Sociological Review* 66 (5): 623-46. https://doi.org/10.2307/3088951.

Heen, Mary. 2009. « Ending Jim Crow Life Insurance Rates ». *Northwestern Journal of Law & Social Policy* 4 (2): 360.

Heller, Douglas. 2015. « High Price of Mandatory Auto Insurance in Predominantly African American Communities · Consumer Federation of America ». Consumer Federation of America. https://consumerfed.org/reports/high-price-of-mandatory-auto-insurance-in-predominantly-african-american-communities/.

Hoffman, Frederick L. 1931. « Cancer and Smoking Habits ». *Annals of Surgery* 93 (1): 50-67.

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

Johnston, Lennox. 1945. « Effects of Tobacco Smoking on Health ». *British Medical Journal* 2 (4411): 98.

Kita-Wojciechowska, Kinga, et Lukasz Kidziński. 2019. « Google Street View Image Predicts Car Accident Risk ». *Central European Economic Journal* 6 (53): 152-63.

Kiviat, Barbara. 2019. « The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores ». *American Sociological Review* 84 (6): 1134-58. https://doi.org/10.1177/0003122419884917.

Kleinberg, Jon, Sendhil Mullainathan, et Manish Raghavan. 2016. « Inherent Trade-Offs in the Fair Determination of Risk Scores », septembre. https://arxiv.org/abs/1609.05807v2.

Knights, D., et T. Vurdubakis. 1993. « Calculations of Risk: Towards an Understanding of Insurance as a Moral and Political Technology ». *Accounting, Organizations and Society* 18 (7): 729-64. https://doi.org/10.1016/0361-3682(93)90050-G.

Kusner, Matt J., et Joshua R. Loftus. 2020. « The Long Road to Fairer Algorithms ». *Nature* 578 (7793): 34-36. https://doi.org/10.1038/d41586-020-00274-3.

Larson, Jeff, Julia Angwin, Lauren Kirchner, et Surya Mattu. 2017. « How We Examined Racial Discrimination in Auto Insurance Prices ». ProPublica. 2017. https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-methodology?token=oXaDaCvsdX3ZY7-YJd8F3L-6fSTJ6BUj.

LeCun, Yann, Yoshua Bengio, et Geoffrey Hinton. 2015. « Deep Learning ». *Nature* 521 (7553): 436-44. https://doi.org/10.1038/nature14539.

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

Lupton, Deborah. 2014. « Self-Tracking Modes: Reflexive Self-Monitoring and Data Practices ». SSRN Scholarly Paper ID 2483549. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2483549.

———. 2016. « The diverse domains of quantified selves: self-tracking modes and dataveillance ». *Economy and Society* 45 (1): 101-22. https://doi.org/10.1080/03085147.2016.1143726.

Lury, Celia, et Sophie Day. 2019. « Algorithmic Personalization as a Mode of Individuation ». *Theory, Culture & Society* 36 (2): 17-37. https://doi.org/10.1177/0263276418818888.

Macnicol, John. 2006. *Age Discrimination: An Historical and Contemporary Analysis*. Cambridge: Cambridge University Press.

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

Mercat-Bruns, Marie. 2020. « Les rapports entre vieillissement et discrimination en droit : une fertilisation croisée utile sur le plan individuel et collectif ». *La Revue des droits de l'homme. Revue du Centre de recherches et d'études sur les droits fondamentaux*, nº 17 (janvier). https://doi.org/10.4000/revdh.8641.

Metz, Jason. 2020. « Sen. Booker's PAID Act Looks To Eliminate Discriminatory Non-Driving Factors In Auto Insurance Pricing ». Forbes Advisor. 5 octobre 2020. https://www.forbes.com/advisor/car-insurance/paid-act/.

Meyers, Gert. 2018. « Behaviour-based personalisation in health insurance: a sociology of a not-yet market ». PhD Thesis, KU Leuven. https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS2087689&context=L&vid=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US&fromSitemap=1.

Meyers, Gert, et Ine Van Hoyweghen. 2018. « Enacting Actuarial Fairness in Insurance: From Fair Discrimination to Behaviour-based Fairness ». *Science*

as Culture 27 (4): 413-38. https://doi.org/10.1080/09505431.2017.1398223.

Miller, G. H., et D. R. Gerstein. 1983. « The Life Expectancy of Nonsmoking Men and Women ». *Public Health Reports (Washington, D.C.: 1974)* 98 (4): 343-49.

Miller, Michael J. 2009. « Disparate Impact and Unfairly Discriminatory Insurance Rates ». *Casualty Actuarial Society E-Forum*, nº Winter 2009. https://www.casact.org/pubs/forum/09wforum/.

Miller, Michael J, et Richard A Smith. 2003. « The Relationship of Credit-Based Insurance Scores to Private Passenger Automobile Insurance Loss Propensity ». EPIC Actuaries, LLC. https://www.progressive.com/content/PDF/shop/EPIC-CreditScores.pdf.

Moor, Liz, et Celia Lury. 2018. « Price and the person: markets, discrimination, and personhood ». *Journal of Cultural Economy* 11 (6): 501-13. https://doi.org/10.1080/17530350.2018.1481878.

Napoletani, D., M. Panza, et D. C. Struppa. 2011. « Agnostic Science. Towards a Philosophy of Data Analysis ». *Foundations of Science* 16 (1): 1-20. https://doi.org/10.1007/s10699-010-9186-7.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. 1st edition. New York: Crown.

Patterson, James T. 1989. *The Dread Disease: Cancer and Modern American Culture*. Cambridge: Harvard University Press.

Pessach, Dana, et Erez Shmueli. 2020. « Algorithmic Fairness », janvier. https://arxiv.org/abs/2001.09784v1.

Prince, Anya E. R., et Daniel Schwarcz. 2019. « Proxy Discrimination in the Age of Artificial Intelligence and Big Data ». *Iowa Law Review* 105: 1257.

Rebert, Lisa, et Ine Van Hoyweghen. 2015. « The Right to Underwrite Gender. The Goods & Services Directive and the Politics of Insurance Pricing ». *Tijdschrift Voor Genderstudies* 18 (4): 413-31.

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

Rudin, Cynthia, et Joanna Radin. 2019. « Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition ». *Harvard Data Science Review* 1 (2). https://doi.org/10.1162/99608f92.5a8a3a3d.

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

Schmeiser, Hato, Tina Störmer, et Joël Wagner. 2014. « Unisex Insurance Pricing: Consumers' Perception and Market Implications ». *The Geneva Papers on Risk and Insurance - Issues and Practice* 39 (2): 322-50. https://doi.org/10.1057/gpp.2013.24.

Shikhare, Shrinivas. 2021. « AI Enabled Next Generation LTC and Life Insurance Underwiring Using Facial Score Model ». In , 19. London. https://insurancedatascience.org/downloads/London2021/Session_4b/Shrinivas_Shikhare.pdf.

Simon, Jonathan. 1988. « The Ideological Effects of Actuarial Practices ». *Law Social Review* 22: 771-800.

Society of Actuaries, (SOA). 1982. « Report of the Task Force on Smoker/NonSmoker Mortality ». Transactions of Society of Actuaries. https://www.soa.org/globalassets/assets/library/research/transactions-

reports-of-mortality-moribidity-and-experience/1980-89/1982/january/TSR8210.pdf.

Swedloff, Rick. 2014. « Risk Classification Big Data (R)Evolution ». *Connecticut Insurance Law Journal* 21 (1): 339-73.

Thiery, Yves, et Caroline Van Schoubroeck. 2006. « Fairness and Equality in Insurance Classification ». *The Geneva Papers on Risk and Insurance - Issues and Practice* 31 (2): 190-211. https://doi.org/10.1057/palgrave.gpp.2510078.

Walters, Michael A. 1981. « Risk Classification Standards ». *Proceedings of the Casualty Actuarial Society* 68: 1-23.

Watson Coleman, Bonnie. 2019. « H.R.3693 - 116th Congress (2019-2020): Prohibit Auto Insurance Discrimination Act ». Legislation. 2019/2020. 11 juillet 2019. https://www.congress.gov/bill/116th-congress/house-bill/3693/.

Wiggins, Benjamin Alan. 2013. « Managing Risk, Managing Race: Racialized Actuarial Science in the United States, 1881-1948 ». Minnesota. http://conservancy.umn.edu/handle/11299/159587.

Williams, Betsy Anne, Catherine F. Brooks, et Yotam Shmargad. 2018. « How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications ». *Journal of Information Policy* 8: 78-115. https://doi.org/10.5325/jinfopoli.8.2018.0078.

Works, Robert. 1977. « Whatever's FAIR—Adequacy, Equity, and the Underwriting Prerogative in Property Insurance Markets ». *Nebraska Law Review* 56 (3): 445-64.

PARI

PROGRAMME DE RECHERCHE
SUR L'APPRÉHENSION DES RISQUES
ET DES INCERTITUDES

PARI, placé sous l'égide de la Fondation Institut Europlace de Finance en partenariat avec l'ENSAE/Excess et Sciences Po, a une double mission de recherche et de diffusion de connaissances.

Elle s'intéresse aux évolutions du secteur de l'assurance qui fait face à une série de ruptures : financière, réglementaire, technologique. Dans ce nouvel environnement, nos anciens outils d'appréhension des risques seront bientôt obsolètes. PARI a ainsi pour objectifs d'identifier leur champ de pertinence et de comprendre leur émergence et leur utilisation.

L'impact de ses travaux se concentre sur trois champs :

- les politiques de régulation prudentielle dans un contexte où Solvabilité 2 bouleverse les mesures de solvabilité et de rentabilité (fin du premier cycle de la chaire);
- les solutions d'assurance, à l'heure où le big data déplace l'assureur vers un rôle préventif, créant des attentes de personnalisation des tarifs et de conseil individualisé ;
- les technologies de data science appliquées à l'assurance, modifiant la conception, l'appréhension et la gestion des risques.

Dans ce cadre, la chaire PARI bénéficie de ressources apportées par Addactis, la CCR, Generali, Groupama, la MGEN et Thélem.

Elle est co-portée par **Pierre François,** chercheur au CNRS, doyen de l'Ecole Doctorale de Sciences Po et **Laurence Barry,** chercheur à Datastorm, la filiale de valorisation de la recherche de l'ENSAE.

PARTENAIRES