Working paper

13

PROJET DE RECHERCHE PARI 2018-2021

LES ENJEUX DU BIG DATA POUR L'ASSURANCE

Pierre François, Laurence Barry

Novembre 2018

PARI

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

PROJET DE RECHERCHE PARI 2018-2021: LES ENJEUX DU BIG DATA POUR L'ASSURANCE

Descriptif détaillé du contexte, des axes de recherche et des champs d'application de l'étude

Pierre François[1]

Laurence Barry[2]

Introduction : contexte et enjeux

L'assurance, un secteur clé d'un point de vue économique et social en pleine évolution

Le secteur de l'assurance recouvre des activités très variées. D'un point de vue technique tout d'abord, il regroupe des domaines touchant de nombreux pans de notre vie quotidienne, tels que l'assurance dommage (*e.g.* multi-risque habitation, les catastrophes naturelles), la responsabilité civile (*e.g.* automobile), la santé ou la prévoyance, ou encore la retraite et l'épargne. Du point de vue de la clientèle ensuite, les assurés étant aussi bien des particuliers que des professionnels, des petites entreprises ou de grands groupes (flottes, risque industriels, assurances collectives, etc.). D'un point de vue commercial enfin, avec des intermédiaires aussi diversifiés que des réseaux salariés, des agents d'assurance, des courtiers, ou encore des prestataires de service.

On retrouve cette diversité dans les formes juridiques des acteurs qui assurent la couverture de ces risques: sociétés anonymes rémunérant un actionnaire, sociétés d'assurance mutuelles relevant du Code des assurances, institutions de prévoyance relevant du Code de la sécurité sociale ou mutuelles 45 relevant du Code de la mutualité. Certaines sont très spécialisées et d'autres fortement diversifiées, que ce soit sous l'angle de l'activité, du périmètre géographique ou encore du champ de clientèle.

La diversité des activités et des acteurs traduit l'importance des mécanismes d'assurance dans le fonctionnement des sociétés industrielles puis post-industrielles pour la gestion des risques qu'elles génèrent. En place notamment dans les assurances publiques et privées, ils sont traditionnellement guidés par des principes de solidarité et des techniques de mutualisation divers.

L'assurance semble toutefois connaître ces dernières années des mutations profondes liées à l'évolution des activités économiques impliquant l'émergence de nouveaux risques, mais aussi à l'arrivée à maturité de technologies qui remettent en cause les principes de solidarité que l'on considérait acquis et les techniques de mutualisation.

L'émergence de nouveaux risques peut se mesurer à deux échelles ;

[1] pierre.francois@sciencespo.fr, Centre de sociologie des organisations, 84 rue de Grenelle, 75007 Paris, Directeur de recherche au CNRS (Sciences Po/CSO), Doyen de l'Ecole Doctorale de Sciences Po, co-porteur de la chaire PARI.

[2] laurence.barry@datastorm.fr , Datastorm, 60 rue Etienne Dolet, 92240 Malakoff, France; co-porteur de la chaire PARI.

- Tout d'abord au niveau macro-économique, voire planétaire : là où Ulrich Beck parlait dans les années 80 et 90, de sociétés de "risque", pour lesquelles cette notion devient omniprésente, la littérature académique envisage l'ère actuelle comme celle de l' « Anthropocène » (Steffen, Crutzen, et McNeill 2007; Schuilenburg et Peeters 2017). La globalisation des risques climatiques et environnementaux s'y accompagne de la reconnaissance de leur difficile mitigation (Bovari, Giraud, et Mc Isaac 2018).
- Au niveau micro-économique ensuite, la révolution digitale de ces dernières décennies transformant les comportements et les rapports sociaux. Portée par la technologie, l'économie du partage par exemple, place l'usage au cœur des échanges, et remet en cause la centralité de la propriété, notion clef du capitalisme classique. Elle implique aussi l'apparition de risques nouveaux, liées à la technologie elle-même, tels que le cyber-risque.

Les technologies du *big data*, qui permettent à la fois la collecte de données de masse et leur traitement au moyen d'algorithmes de plus en plus performants, achèvent de bouleverser le paysage de l'assurance. En effet, avec une granularité de plus en plus fine des informations disponibles, il devient possible aujourd'hui d'appliquer des statistiques au niveau de l'individu, pratique inenvisageable dans la période précédente.

La « révolution » du big data …

La notion de « *big data* » est délicate à cerner. Si l'on admet avec Ollion et Bollaert (2015) que les données de masse désignent en première approximation « ces vastes ensembles de données, dont l'existence comme le traitement rapide ont été rendus possibles par une série de changements technologiques », on voit qu'elles doivent se décrire sur deux dimensions, celle de leur nature (en quoi consistent ces données ?) d'abord, celle de leur traitement (qu'en fait-on ?) ensuite, et que leur émergence doit se saisir dans un moment singulier, celui des bouleversements technologiques qui présideraient à leur émergence. Nous distinguerons ces deux dimensions en nous efforçant d'apprécier la nouveauté – radicale ou relative – des interrogations qu'elles soulèvent.

- *Quelles données ?*

La période contemporaine n'est pas la première de l'histoire des sociétés contemporaines à voir émerger des ensembles de données qui ouvrent, a priori, de nouvelles perspectives à la compréhension des sociétés et des pratiques individuelles : entre la fin du XVIIIe siècle et la première moitié du XXe siècle se construisent ainsi les appareils de collecte et de traitement de la statistique publique dans la plupart des grands pays occidentaux (Hacking 1990; Desrosières 1993), tandis que les années d'après-guerre voient se démultiplier les initiatives visant à rassembler des données sur la base d'enquêtes par échantillon, à des fins commerciales (Cochoy, 1999 ; Blondiaux, 1998) ou savantes (Calhoun et van Antwerpen, 2007 ; Abbott et Sparrow, 2005). Une manière classique de caractériser la nouveauté du *big data* tiendrait à leurs caractéristiques, synthétiquement résumées par les « 3 V » :

- Leur *volume*, tout d'abord : la quantité des données individuelles disponibles serait sans commune mesure avec celle qui prévalait jusqu'à la fin des années 1990, grâce aux téléphones portables qui collectent des données en continu, mais aussi aux capteurs placés dans nos voitures ou aux bracelets connectés qui enregistrent les pulsions cardiaques ou décomptent nos pas, sans compter les pratiques de navigation sur internet (sites visités, *clicks* et *likes*, vidéos ou musiques consommées, *posts* sur les réseaux sociaux…), qui alimentent un puits apparemment sans fond de données.
- Leur *variété*, ensuite : ces données sont d'une nature extrêmement hétérogène (textes, images, vidéos, etc.), conséquence de l'extension a priori infinie de la sphère du

quantifiable qui consiste, selon la formule de Desrosières, à « exprimer et faire exister sous une forme numérique ce qui, auparavant, était exprimé par des mots et non par des nombres » (Desrosieres 2008b, p. 10-11).

- Leur *vélocité*, enfin, qui renvoie à la fois à l'instantanéité de leur collecte (les informations collectées peuvent l'être en temps réel, elles ne sont plus frappées de la caducité propre aux formes classiques d'enquête) et à la rapidité des traitements dont elles peuvent faire l'objet – nous y reviendrons plus loin.

Cette première caractérisation des données relevant du *big data* permet de souligner l'une de leurs principales caractéristiques : à la différence des informations recueillies dans les enquêtes des instituts de statistiques publiques, des bureaux de marketings ou des entreprises de sondage d'opinions, elles ne sont pas produites avant tout à des fins de connaissance – il s'agit, selon les termes de Groves, de données organiques, *i.e.* de données qui émergent des pratiques elles-mêmes et de l'enregistrement qu'en font les dispositifs techniques sur lesquelles elles s'appuient : « *Internet search engines build data sets with every entry, Twitter generates tweet data continuously, traffic cameras digitally count cars, scanners record purchases, RFID's signal the presence of packages and equipment, and internet sites capture and store mouse clicks. Collectively, the society is assembling data on massive amounts of its behaviors.* » (Groves 2011, 868). Le caractère organique des données du *big data* implique qu'elles ne sont pas *a priori* organisées pour faire l'objet de traitement élaborés : il s'agit très souvent de données intrinsèquement pauvres (on sait que Robert a cliqué trois fois sur ce lien, mais on ne sait rien de Robert – sauf à pouvoir mettre en regard ces trois clics avec d'autres de ses pratiques et avec ses propriétés sociales) et désordonnées (Cukier et Mayer-Schoenberger (2013) disent ainsi des données du *big data* qu'elles sont « *messy* »).

La pauvreté et le désordre de ces données ne sont, à dire vrai, pas spécifiques au *big data* – et les chercheurs en sciences sociales disposent d'outils et de savoir-faire qui permettent de travailler avec un matériau qui n'était pas *a priori* destiné à produire de la connaissance : depuis la moitié du XIXe siècle, les historiens ont entrepris de définir les règles élémentaires de la mise en critique et du recoupement des archives qu'ils dépouillent (voir, par exemple, les recommandations canoniques de Langlois et Seignobos (1898)), ainsi que de la constitution de séries statistiques robustes sur la base d'un matériau au moins aussi « *messy* » que celui qu'offrent les données de masse (Lemercier et Zalc, 2008). De ce point de vue, les défis – bien réels – que posent ces données sont moins neufs qu'on ne le dit souvent, et ils tiennent au moins autant à l'extension de la population des acteurs qui veulent en faire usage qu'à celles des données disponibles. Rapprocher les enjeux d'une organisation et d'un usage rigoureux de ces données de la pratique ancienne et contrôlé du rapport à l'archive n'annule pas les difficultés mais les déplace, en même temps que s'identifient un ensemble d'exigences et de techniques que les praticiens du *big data*, quels qu'ils fussent, pourraient opportunément endosser.

Un reproche traditionnellement adressé à la notion de *big data* et à la caractérisation classique (celle des « 3V ») qui peut en être faite tient à leur extrême hétérogénéité : quoi de commun, en effet, entre l'enregistrement des pas sur un plancher équipé de capteurs, la numérisation de dizaine de milliers d'ouvrages, la passation de centaine de milliers de questionnaires en lignes, la mobilisation de données que les organisations recueillent sur leurs employés ou leurs clients ? Le *big data* s'est imposé comme un terme étendard – mais les termes étendards sont rarement les mieux placés pour clarifier les enjeux qui leur sont attachés. Une première distinction, avancée par Cukier et Mayer-Schoenberger (2013), permet de distinguer entre la *datafication* et la *digitalisation*. La *datafication* désigne la capacité à transformer en données des dimensions de nos existences qui jusque-là échappaient à toute forme de quantification ou de mise en série. La localisation, par exemple, a été « datafiée » avec la formalisation des notions de latitude et de longitude, d'abord, puis avec la mobilisation des GPS qui permet de les enregistrer de manière systématique. La *datafication* n'impose pas nécessairement la mobilisation de données ou de techniques digitales ou numériques : les bases de données des historiens dont nous parlions plus

tôt peuvent consister en une datafication des données biographiques des individus, sans que ces données soient mises en ligne ou traitées numériquement pour être exploitées. La *digitalisation* désigne quant à elle un autre processus, à bien des égards plus spécifique, en ce qu'elle consiste à transformer des matériaux analogiques (des textes, des films, des images) et à les digitaliser en les transformant en une série de zéro et de un que peuvent lire les ordinateurs.

La distinction entre datafication et digitalisation, que l'on peut éventuellement entendre sur un mode séquentiel (la digitalisation précédant la datafication), invite à introduire des segmentations au sein de la masse trop hétérogène des données du *big data*. On peut pour cela suivre les propositions d'Ollion et Bollaert (2015) qui font également jouer à la numérisation un rôle central. Ils proposent ainsi de distinguer :

- *Les données de l'internet*, qui désignent « les informations qui sont collectées, ou auxquelles on accède, via le net » (Ollion et Bollaert, 2015, p. 300). Il faut à cet égard distinguer les informations relatives aux pratiques en ligne (quels sites visite-t-on, par exemple) et celles qui renvoient, de manière plus ou moins précise, à des pratiques *offline* (par exemple, la fréquentation d'un stade pour lequel on achète des billets en ligne).
- *Les données produites par les organisations* (entreprises, association, administration) dans le cadre de leur fonctionnement. Depuis très longtemps les organisations produisent, au cours de leur activité, des informations relatives, par exemple, à leur clientèle, à leurs salariés, au fonctionnement de leurs services. Ces données supposent la mise en place de systèmes d'informations internes souvent lents à se construire (Chiapello et Gilbert, 2009). De ce point de vue, la mobilisation en leur sein de ce qu'elles nomment le « *big data* » ne constitue qu'une étape supplémentaire dans un processus séculaire, qui ouvre peut-être de nouvelles perspectives mais qui, assurément, ne va pas de soi et impose de substantielles réorganisations de la gestion de ces systèmes d'informations : un responsable de la mise en regard des différentes sources d'informations disponibles dans un grand groupe d'assurance, évoquant l'hétérogénéité des formats de saisie des dates de naissance des assurés selon qu'ils souscrivent tel ou tel type de contrats, énonçait ainsi qu' « avant de faire du *big data*, il va falloir faire du *data* ».

  Ces données, quoiqu'il en soit, sont propriétaires, et elles ne sont à ce titre que difficilement accessibles. Celles de ces données qui sont mises en ligne relèvent des *open data*. Sur les portails qui les mettent à la disposition de leurs utilisateurs potentiels se trouvent des données d'une grande hétérogénéité. Ces *open data* sont très inégalement nourries par les différents types d'organisation : les administrations les alimentent beaucoup plus que les entreprises ou les associations.
- *Les archives numérisées* : il s'agit de données qui au départ ne sont pas en format numérique (on parle de données « non nativement numériques »), mais qui sont converties en format numérique. Certaines entreprises très spectaculaires (des millions de documents non numériques ont été numérisés par Google, tandis que les *internet archives* de la Bibliothèque du Congrès rassemblent plus de 150.000 documents) ont pu nourrir des programmes de recherche comme le courant *culturomics* (Michel et al., 2011).
- *Les questionnaires passés en ligne* : on est ici, *a priori*, proche d'une pratique canonique des sciences sociales, *i.e.* la passation de questionnaires, qui devient à bien des égards plus simple que lorsqu'il fallait les administrer en face-à-face ou au téléphone (le dispositif, très coûteux, n'était qu'employé que lorsque l'enquête bénéficiait d'un soutien institutionnel important) ou encore par voie postale (moins onéreuse, l'enquête

souffrait d'un taux de réponse souvent très faible). Il est désormais possible d'interroger une population beaucoup plus nombreuse, comme dans le cas du *Great British Class Survey* qui a recueilli des informations auprès de plus de 150.000 répondants (Savage et *al.*, 2013). Toute la difficulté est alors de contrôler l'échantillon et de maîtriser les biais qui ne manquent pas de s'y faire jour (Dilman et *al.*, 2014).

On voit que le « *big data* » désignent en réalité des données de nature très hétérogène, et que les enjeux qui leur sont attachés diffèrent très profondément selon que l'on parle de tel ou tel type de données. Cette hétérogénéité des données est comme démultipliée par celle des outils statistiques ou mathématiques qui peuvent leur être appliqués.

- ***Quels outils ?***

Tout comme les données n'ont pas attendu l'ère du *big data* pour être produites et mises en série, les outils mobilisés pour en rendre compte ont eux aussi une histoire séculaire. L'une des hypothèses qui accompagne la réflexion sur le *big data* consiste à avancer que les outils des disciplines traditionnelles seraient rendus caducs par le changement d'échelle ou de nature qui accompagne l'avènement de ces nouvelles données – soit qu'ils ne soient plus adaptés pour décrire adéquatement ces vastes jeux de données, soit qu'ils soient beaucoup moins performants que d'autres outils développés sur d'autres territoires mais que l'on pourrait maintenant déplacer avec succès. Disons-le d'emblée : les pratiques sont sur ce point très loin d'être stabilisées, et l'on peine à savoir quelles hybridations sont susceptibles de se faire jour entre les outils traditionnels des sciences sociales, ceux avancés par les *computer sciences* ou par les transfuges de la physique. Sur ce point, autrement dit, l'enjeu est double : il s'agit autant de recenser les propositions qui aujourd'hui se font jour – dans des espaces disciplinaires et sociaux souvent très éloignés – que de mettre au jour les attendus normatifs qui sous-tendent ces modèles.

Evoquons d'abord les modèles traditionnels des sciences sociales et les déplacements que l'avènement de larges jeux de données sont susceptibles d'impliquer. On les distribue, classiquement, en grandes familles dont la maîtrise est plus ou moins systématique selon les pays ou les traditions disciplinaires : statistiques descriptives, statistiques inférentielles, analyse géométrique, analyse de réseaux, analyse de séquences… La question de savoir si ces outils sont pertinents pour rendre compte de vastes corpus de données fait l'objet de débats qui ne seront sans doute pas tranchés avant plusieurs années (pour une synthèse provisoire, voir Hampton, 2017). Contentons-nous à ce stade d'évoquer deux discussions qui ont trait, plus particulièrement, à la place des raisonnements statistiques inférentiels pour décrire de très vastes populations ou, au contraire, pour avancer des raisonnements à l'échelle des individus.

- *Des données pour l'infiniment grand, ou a-t-on encore besoin de raisonnements inférentiels* ? Les outils statistiques couramment utilisés en sciences sociales ont été développés pour inférer, à partir de l'étude d'une fraction d'une population (un « échantillon »), des liaisons entre des variables susceptibles de se faire jour au sein d'une population mère que l'on ne peut étudier directement, et pour apprécier la fiabilité de cette inférence (en se demandant par exemple quelle est la probabilité d'observer une liaison entre les deux variables au sein de l'échantillon, alors que cette liaison n'existe pas dans la population mère). Ces outils ont été d'une très grande utilité tant que les conditions d'enquête interdisaient d'interroger une population très étendue – ils permettent en effet de procéder à des inférences de l'échantillon sur la population mère avec des échantillons d'un millier d'individus interrogés.

  Dès lors que les modalités de passation des questionnaires ou de recueil des données permettent de travailler sur des populations beaucoup plus importantes – voire de prétendre travailler sur des populations complètes – ces outils sont-ils encore pertinents ? Saporta (2006) montre par exemple que si l'on dispose d'un million

d'observations, alors l'hypothèse d'indépendance de deux variables numériques sera rejetée au seuil de 5% dès que leur coefficient linéaire dépassera, en valeur absolue, la valeur de 0.002 – ce qui est tout simplement dépourvu de sens : dès lors que les observations sont très nombreuses, les *p-values* sont systématiquement très faibles et toutes les liaisons sont significatives.

Moins utiles quand les observations sont très nombreuses, les outils de la statistique inférentielle sont aussi d'un usage plus délicat, voire inappropriés : pour pouvoir inférer rigoureusement de l'échantillon à la population mère, la constitution de l'échantillon obéit à des règles strictes que la passation de questionnaires ou la collecte de données en ligne rend le plus souvent impossible à respecter. Dès lors, le recours à d'autres outils mathématiques, moins fréquemment mobilisés par les sciences sociales, peut s'avérer plus opportun.

- *Des données pour l'infiniment petit, ou peut-on prédire le comportement individuel ?* L'usage classique de la statistique inférentielle consiste, à partir de données collectées au niveau individuel, à étudier des phénomènes collectifs et à rendre compte d'un autre niveau de réalité, celui de la population. Les données qu'elle utilise sont organisées en tableau, chaque ligne représentant un individu et chaque colonne une des variables de l'analyse. Or, pour reprendre l'image de Desrosières (2014), si la constitution du tableau est horizontale, son analyse est verticale : ce que l'on cherche à comprendre ce sont des relations entre les variables. « Dans le modèle linéaire de la régression, écrit-il, (…) les sujets des verbes, et donc des actions, ne sont plus des personnes ou des groupes sociaux, mais des variables, entités nouvelles, issues d'une série de conventions d'équivalence, de taxinomies, de codages, d'évaluations selon des grilles diverses. Les personnes sont décomposées en items, qui sont eux-mêmes recomposés en variables. L'opérateur est le tableau croisant en lignes des personnes (ou toutes sortes d'autres êtres, individus ou groupes) et, en colonnes, des items codés de façon standardisée sur chacun de ces êtres. Dans le premier monde, ce tableau est lu en ligne, les individus ou les groupes sont les sujets des verbes. On y raconte des histoires. Dans le second monde, celui de la statistique, le regard a tourné à 90 degrés ; le tableau est lu en colonne, les variables sont devenues les acteurs du théâtre. Elles sont désormais les sujets des verbes, elles entrent en relation les unes avec les autres, s'expliquent, sont corrélées positivement ou négativement » (Desrosières, 2014, p. 169). Si l'on conserve l'image du tableau statistique, on peut avancer que le propre du *big data* est de changer l'équilibre ancien des rapports de hauteur et de largeur du tableau : le tableau du statisticien comportait quelques dizaines de variables et un nombre potentiellement très important de lignes. Le traitement statistique des variables était rendu possible par la loi des grands nombres (le grand nombre d'observation d'occurrence de ces variables). Avec le *big data*, on se trouve dans une situation où le tableau devient aussi large que long ; on dispose d'un très grand nombre de données par ligne, c'est-à-dire par individu observé, ce qui rend possible une analyse statistique de l'individu ; ce dernier devient traitable *comme une variable* et l'on peut calculer des corrélations entre individus, ce qui était impensable dans une logique statistique classique (Barry, à paraitre).

Les débats relatifs à l'usage de la statistique inférentielle pour analyser les données du *big data* sont loin d'être clos – et les modèles inférentiels ne recouvrent pas, par ailleurs, l'ensemble des outils mathématiques sur lesquels s'appuient les sciences sociales. L'enjeu n'est pour nous ni de présenter exhaustivement ces débats ni (évidemment !) de les trancher, mais plus simplement d'en donner un aperçu. Il est aussi de faire état des entreprises visant à leur trouver des alternatives.

Avec l'arrivée de jeux de données massives se sont en effet développées des propositions visant à substituer aux outils (et, implicitement, aux savoirs) des sciences sociales de nouveaux modèles. C'est l'ambition explicite de certaines propositions, qui parfois s'agencent dans des constructions théoriques explicites (comme la *social physics* (Pentland, 2014), la *Generative social science* (Epstein, 2006) ou la *Web Science* (Hendler et Al, 2008)), et parfois se déploient dans des entreprises collectives, transversales et hétérogènes, comme les *computer social sciences* (Conte et *al.*, 2010) ou la *science of networks* (Watts, 2004) – pour ne prendre que quelques exemples. C'est aussi l'enjeu de la mobilisation d'outils mathématiques ignorées jusque-là des sciences sociales et de leur application à des jeux de données qui étaient jusque- là jugés trop imparfaits (*i.e.,* le plus souvent, trop réduits) pour produire un savoir inédit et robuste sur les phénomènes collectifs ou individuels que les données de masse permettent désormais d'approcher. Ces démarches, dont beaucoup sont récentes et très ambitieuses, ont souvent en commun de laisser de côté les résultats et les démarches des sciences sociales traditionnelles – ce qui, on s'en doute, ne manquent pas d'agacer ceux qui voient arriver sur leur terre ces nouveaux venus enthousiastes et révolutionnaires, comme en témoignent les remarques de Duncan Watts lorsqu'il souligne que les propositions des physiciens, des biologistes ou des mathématiciens qui investissent la « nouvelle » science des réseaux innovent souvent assez peu (Watts, 2004). Beaucoup de ces démarches doivent par ailleurs faire la preuve de leur portée empirique : le programme de recherche d'Epstein, extrêmement ambitieux, a bien été testé sur des données tirées de *twitter* par Smith et Broniatowsky (2016), mais sa portée reste encore très largement à démontrer. Les outils alternatifs qui émergent et qui pourraient être utilisés pour traiter les vastes jeux de données désormais disponibles sont donc à la fois très fortement segmentés, souvent assez techniques et malaisés d'approche, et doivent encore faire la preuve de leur efficacité : c'est peu dire, dans ces conditions, que sous ce jour, le champ du *big data* est encore très largement en construction.

De ce foisonnement émergent enfin de nouveaux modes de traitement *a priori* particulièrement adaptés à l'extension des nouveaux jeux de données, qui se regroupent sous l'appellation générique du *machine learning*. Ce vocable est lui aussi assez flottant. Il recouvre parfois, de façon large, l'ensemble des modèles faisant appel à des algorithmes - entendus comme des calculs itératifs visant à minimiser l'erreur, c'est-à-dire aussi aujourd'hui la totalité des modèles statistiques classiques (et de ce point de vue, une régression « classique » est aussi du *machine learning*). Au cours des vingt dernières années, l'accroissement des capacités de calcul a cependant permis de mettre en application de nouveaux algorithmes, souvent très complexes, qui sont ceux visés par une définition plus restreinte du *machine learning*. Ces méthodes, dont les principes théoriques ont été posés à l'intersection des mathématiques et de l'informatique entre la fin des années 1940 et le début des années 1960, ont connu un essor à partir des années 1990 et 2000 avec, d'une part, un réinvestissement théorique appuyé, notamment, sur les mathématiques et les statistiques, et, d'autre part, sur l'accès à des données beaucoup plus volumineuses et à une puissance de calcul inédite, tous deux nécessaires à la mise en œuvre de ces modèles. Ils constituent aujourd'hui l'une des pistes qu'explorent simultanément les chercheurs en sciences sociales et les spécialistes de *computer science*, pour préciser la nature des raisonnements qu'ils mettent en œuvre au regard des outils statistiques traditionnels, et pour apprécier les propriétés des résultats qu'ils permettent de mettre au jour.

Quant à la nature des raisonnements déployés dans ces modèles, ils reposent sur des bases théoriques très largement disjointes de ceux de la statistique inférentielle (pour une présentation de l'une des méthodes aujourd'hui en vogue, celle du *deep learning*, voir, entre autres, LeCun et *al.*, 2015). Fawcett et Hardin saisissent en une image la parenté intuitive et les indéniables différences qui se font jour entre les deux types de raisonnement : *« [Statistics and machine learning] are like two pairs of old men sitting in a park playing two different board games. Both games use the same type of board and the same set of pieces, but each plays by different rules and has a different goal because the games are fundamentally*

different. Each pair looks at the other's board with bemusement and thinks they're not very good at the game"[3]. Une manière synthétique de caractériser les principales différences qui se font jour entre les deux types de raisonnement consiste à souligner qu'alors que les méthodes statistiques s'appuient sur des échantillons de taille restreinte et des modèles eux-mêmes appuyés sur des hypothèses très spécifiques, les modèles du *machine learning* émergent d'un apprentissage qui, pour être efficace, supposent de pouvoir être mis en œuvre sur des jeux de données très importants. Ces modèles ont ainsi plusieurs caractéristiques qui les différencient des statistiques classiques :

- L'approche classique consistait à poser un modèle *a priori* et à déterminer ses paramètres. Mais, pour Breiman (2001), les conclusions tirées de ce type d'approche portent sur le modèle lui-même et non sur le phénomène que l'on cherche à modéliser à travers lui. Les méthodes spécifiques au *big data* ne posent aucune formule *a priori* et essaient de « coller » aux observations sans présumer d'une formule.
- Le modèle dominant des statistiques classiques est la régression linéaire, *i.e.* la mise en évidence de relations (linéaires) entre des variables, supposées normales. Même si cette hypothèse est rarement vérifiée en pratique et bien qu'il soit admis de longue date que cette représentation « linéaire » de la réalité est extrêmement restrictive (Abbott 1988), elle a l'avantage de se prêter à une interprétation simple (simpliste ?) des phénomènes.
- Dans les nouveaux modèles, c'est à la fois la normalité des variables, la linéarité de leurs relations et la simplicité de la formule qui sont abandonnées : cette approche se veut plus scientifique au sens où elle serait mieux ajustée à la réalité que l'on cherche à modéliser. Pour citer Breiman : « *the approach is that nature produces data in a black box whose insides are complex, mysterious, and, at least, partly unknown*" (Breiman 2001, 205).
- C'est plus fondamentalement le sens de la modélisation qui a changé : les outils statistiques sont traditionnellement engagés dans une démarche sinon explicative, au moins interprétative, comme la recherche des causes sous-jacentes aux phénomènes. Les méthodes tirées du *machine learning* ont avant tout pour ambition de décrire (et notamment de classer) ou de prédire. Pour reprendre une expression célèbre, les méthodes du *big data*, très efficaces à reproduire les observations, semblent annoncer « la fin des théories » (Anderson 2008), comme si une nouvelle forme de science, purement prédictive, était en train de se mettre en place.
- Cette dernière proposition – qui est par ailleurs toujours fort discutée – nous amène vers un autre point de contraste entre méthode statistique et *machine learning*, qui porte sur la nature des résultats produits par les deux démarches. En comparant les résultats d'une analyse en composante principale classique et d'une méthode appelée t-SNE (pour (*t-distributed stochastic neighborhood embedding*, Van der Maaten & Hinton, 2008) Ollion et Bolaert (2015) montrent que la première décrit les espaces des individus sous une forme aisément interprétables – mais perd au passage de nombreuses informations, tandis que la seconde, plus fine dans sa portée descriptive, est nettement plus ardue à interpréter. De même, notent les auteurs, « là où les coefficients d'une régression linéaire sont immédiatement interprétables et répondent à des questions précises – la variable x contribue-t-elle significativement aux variations de la variable y ? Quelle augmentation moyenne une incrémentation de x entraîne-t-elle sur y, toutes choses égales par ailleurs ? – ce n'est pas le cas des paramètres d'une forêt aléatoire » (Ollion et Bolaert, 2015, 305).
- Face aux critiques souvent avancées devant cette dimension boîte noire, se développent ces dernières années diverses initiatives pour rendre les modèles plus

[3] *Machine Learning vs. Statistics. The Texas death match of data science*. August 10th, 2017 – https://www.svds.com/machine-learning-vs-statistics/, consulté le 25 mai 2018.

explicites : une nouvelle forme d'intelligence artificielle dite « XAI », vise par exemple à faire tourner, en parallèle au modèle principal un deuxième réseau neuronal, de type sémantique, pour rendre compte des paramètres et des variables impliqués dans chaque neurone (Kuang 2017). Alors que l'entrée en vigueur du RGPD semble imposer ce type de techniques explicatives, leur efficacité (voire leur possibilité) reste cependant à démontrer (Bornstein 2016 ; Wallace 2017).

Les débats autour de la lisibilité des modèles spécifiques au *big data* ne vient pas seulement de leur complexité, mais aussi de l'opacité de leur usage : ce qui était jusqu'il y a peu limité au *credit scoring* semble en effet se généraliser à tous les aspects de la vie quotidienne. Comme le dit Pasquale, alors que les individus deviennent de plus en plus transparents aux organismes qui collectent et traitent les données, les usages qui en sont faits restent très mal connus (Pasquale 2015).

…Et ses implications pour l'assurance

L'une des conséquences de la généralisation de ces nouvelles technologies est la possibilité de modifier de façon significative la segmentation des risques et par là les modèles classiques de tarification. Partant d'une logique de mutualisation sous-tendue par un idéal de solidarité entre les assurés, la segmentation de plus en plus granulaire - rendue possible par l'accumulation des données de masse et des algorithmes qui les accompagnent - laisse envisager une assurance personnalisée à outrance où la solidarité, voire la mutualisation, semblent avoir disparu. Cette capacité pose de façon renouvelée la question des principes de la tarification des risques.

La gestion des risques, dans les techniques d'assurance traditionnelles implique en effet leur mutualisation : pour pouvoir accéder à une régularité statistique, l'assureur s'appuie sur l'hétérogénéité de situation ex-post de ses assurés et la loi des grands nombres. Le risque est géré sur la base de statistiques collectées sur un grand groupe d'individus devenus ainsi solidaires. La prédiction au niveau individuelle en revanche est techniquement et conceptuellement impossible, et réputée relever de l'aléa.

Cette impossibilité semble levée aujourd'hui par la capacité de collecter des informations de plus en plus granulaires sur chaque individu. Le *big data* alimente ainsi le fantasme d'une prédiction individuelle et d'un tarif « ajusté » au risque personnel, exact, supposé connu ou connaissable, de chacun. Mais derrière ce fantasme, c'est le principe de mutualisation, voire la possibilité d'assurance qui sont ébranlés.

*

Quels pourraient-être les principes d'une mutualisation « intelligente » dans ce contexte émergent du *big data*, tel est l'objet des recherches proposées ici. Nous évoquerons successivement des interrogations relatives à la production du *big data* (i.e. de ses données et de ses modèles) et des questionnements sur ses implications, pour les individus, les organisations ou les régimes de solidarité collective, autour l'émergence potentielle d'un nouveau paradigme.

La fabrique du big data

Nous rappelions plus haut les incertitudes qui ont trait à l'identification de ce qu'est le *big data* et les difficultés que l'on peut rencontrer à en cerner le contenu et les contours. Une manière, sinon de lever cette difficulté, du moins de la contourner temporairement, consiste à renvoyer l'interrogation aux acteurs : que nous disent-ils de ce qu'est le *big data* ? L'enjeu est moins ici de procéder au relevé cadastral des 1001 définitions indigènes que de faire de cette question une porte d'entrée pour nous intéresser à la production des données et des formes de traitement que les acteurs mettent sous le terme de *big data*. Il ressort en effet des définitions exploratoires que nous présentions plus haut que le *big data* est tout sauf un champ stabilisé, et moins encore un donné : pour que des jeux de données massifs soient mobilisables, il faut que ces données soient produites et assemblées, que des bases de données disparates puissent être appariées – il faut en un mot un travail de production des données dont nous souhaitons analyser les déclinaisons organisationnelles au sein, plus particulièrement, des compagnies d'assurance ; et pour que ces jeux de données soient exploités, il faut que les outils mathématiques qui permettent de les analyser soient eux aussi produits, et qu'ils soient appropriés par des communautés professionnelles souvent très éloignées de leur espace de production initial. Ce sont ces deux volets – travail de production des données et processus d'appropriation des modèles – que nous souhaitons mettre au cœur de ce volet de recherche.

La production des données

L'appareil statistique qui est aujourd'hui au cœur des sociétés contemporaines a été lent à se construire : la mise en place de son infrastructure, étudiée en profondeur dans les cas français (Desrosières 1993), allemand (Labbé, 2008), italien (Patriarca, 1996) ou américain (Anderson, 1988), a supposé des investissements considérables et des choix politiques souvent extrêmement lourds – bien que souvent implicites. La mise en place des systèmes d'information au sein des entreprises privées a grossièrement coïncidé avec le développement des systèmes de statistique publique et, de même que la statistique publique est indissociable de la mise en place des Etats modernes (Desrosières 1993), de même le développement de systèmes de quantifications propres aux entreprises privées est lui aussi indissociable du développement de bureaucraties rationalisées portées vers la maximisation du profit. Dans son *Economie et Société*, Weber (1995) fait ainsi du développement de la comptabilité en partie double l'un des plus sûrs symptômes – en même temps que l'une des causes principales – de l'avènement du capitalisme moderne : parce que seule la comptabilité en partie double est susceptible de rendre possible la mise en œuvre d'un calcul systématique et informée, la recherche du profit pour lui-même, note Weber, trouve dans le développement de ce qu'il nomme le « compte capital », l'une de ses plus nécessaires conditions de possibilités. De la même manière, Chandler (1988) fait coïncider la naissance de la grande firme moderne, typique selon lui du capitalisme contemporain, de la mise en place des systèmes d'information internes aux entreprises qui rendent possible le contrôler des grandes bureaucraties productives.

Cette lecture très fonctionnelle du développement des entreprises modernes et des systèmes d'informations qu'elles mettent sur pied pour en rendre possible la gestion a été assez profondément remise en cause par des travaux qui soulignent que les différents systèmes de comptabilité sont, longtemps, moins des outils de mise en calcul systématique du monde qu'un dispositif mnémonique plus modeste (mais non moins crucial), et qu'il s'agit souvent moins, pour ceux qui les mobilisent, de les utiliser pour calculer que pour manifester leur sérieux, leur bonne foi et leur modernité auprès de leurs partenaires potentiels. Et si ces outils s'imposent et se généralisent, ce n'est pas tant parce que l'évidence de leur efficacité s'imposerait aux yeux de l'ensemble des acteurs économiques, mais parce que les Etats les imposent pour nourrir leur

propre système de comptabilité publique, et aussi pour assoir sur des règles précises leur politique fiscale (Carruthers et Espeland, 1991 ; Gervais, 1992). Qu'il obéisse à des logiques fonctionnelles ou symbolico-politiques, le développement des grandes entreprises modernes est de toutes façons inséparable du développement de systèmes d'informations dont la mise en place a supposé que soient consentis des investissements considérables.

Replacé dans la ligne de fuite de cette histoire longue, le développement contemporain du *big data* au sein des entreprises constitue bien plus le prolongement d'une trajectoire séculaire qu'une rupture brutale ouvrant sur des interrogations inédites. C'est dans cet horizon de long terme qu'il conviendra de replacer les enjeux touchant la construction des larges jeux de données susceptibles d'être mobilisés par les entreprises d'assurance. Dans l'éventail hétérogène de ce que recouvre le *big data* et que nous rappelions plus haut, les données mobilisables par les compagnies d'assurance relèvent avant tout des données produites par les organisations. De ce point de vue, l'enjeu est moins de comprendre comment seraient produites de nouvelles données que de décrire la manière dont des données pré-existantes – et *a priori* non destinées à être engagées dans des analyses systématiques – sont susceptibles d'être mobilisées à des fins renouvelées. La constitution de ces larges bases de données est très loin d'aller de soi. Pour pouvoir faire l'objet de traitements systématiques et robustes, les données doivent en effet faire l'objet d'un travail systématique de codage et d'appariement, qui n'est pas placé au principe de la construction de bases de données dont le propos est avant tout commercial, juridique ou organisationnel. On peut schématiquement distinguer quatre types d'enjeux qui seront placés au cœur de nos interrogations :

- Le premier enjeu, le plus évident peut-être, est aussi celui qui a été le moins systématiquement exploré par la littérature récente – même si, comme le soulignent Berman et Hirschman (2018), la plupart des travaux portant sur la quantification évoquent la manière dont les données quantifiées sont produites : il consiste à mettre au jour les conditions organisationnelles de possibilité de ces bases de données. Une grande part des jeux de données susceptibles d'être mobilisées par les grandes entreprises d'assurance repose, nous venons de le dire, sur des données déjà collectées, mais le plus souvent distribuées en des points épars de l'organisation – et leur rassemblement est très loin d'aller de soi. Des difficultés techniques peuvent évidemment se faire jour : les formats de données du service A sont très rarement ceux des données du service B, et cette hétérogénéité qui ne soulevait aucune difficulté lorsque les activités opérationnelles des deux services se déployaient en parallèle peut s'avérer rédhibitoire lorsqu'il s'agit d'apparier leur base… Mais les travaux sur les pratiques de quantification montrent également que les obstacles techniques ne sont le plus souvent que le prélude à des difficultés politiques – au point que suivre la circulation d'une information et repérer les obstacles qu'on lui oppose constitue souvent un excellent point d'entrée pour repérer les lignes de fracture et les luttes de pouvoir qui peuvent se faire au sein d'une bureaucratie (Voir notamment les travaux rassemblés par Karasti et al., 2016, qui présentent ces enjeux à partir de la notion « d'infrastructure du savoir »). Reconstituer, comme nous entendons le faire, les voies de construction des bases de données internes aux entreprises d'assurance, ne nous permettra pas seulement de dénaturaliser l'évidente disponibilité des *big data* que postulent parfois, pour s'en effrayer ou s'en féliciter, ceux qui en décrivent l'avènement prochain. Nous pourrons également mettre au jour les difficultés proprement organisationnelles attachées à la mise en place de ces jeux de données, en décrivant les segments de l'organisation qui en tirent profit et ceux qui au contraire en pâtissent, en mettant au jour les conflits que peut engendrer cette mise en place et en analysant les modalités de leur règlement.

- La construction de ces bases de données s'appuie sur des professionnels dont les expertises peuvent être très hétérogènes – commerciales, juridiques, actuarielles, etc. – et la nature de ces expertises n'est évidemment pas indifférente. Les travaux récents, qui placent souvent l'expertise au cœur de leur démarche, soulignent ainsi que parce que les indicateurs quantifiés sont produits par des communautés d'experts en lutte les unes avec les autres, ils reflètent autant des rapports de force que des options épistémiques – options épistémiques qui, par ailleurs, ne sont évidemment pas neutres : proposer telle ou telle cartographie du monde est susceptible de favoriser plus ou moins telle ou telle communauté, ne serait-ce qu'en offrant des prises à son action, ou en les lui refusant au contraire (voir, par exemple, Merry, 2016 ; Barman, 2016). L'emprise de l'expertise est par ailleurs au principe de l'inertie qui est au cœur de ces pratiques de cartographie quantifiées : parce qu'elle engage un savoir ésotérique, lent à s'élaborer et à se transmettre, parce qu'elle structure les formes des bases de données, les variables que l'on y trouve et les indicateurs que l'on en peut tirer, l'expertise contribue à limiter le potentiel d'innovation attachée à la construction et aux usages des bases de données. De ce point de vue, l'innovation attendue des *big data* pourrait venir au moins autant de la mobilisation et de l'extension de nouvelles données que de la mise en regard et de la confrontation de formes d'expertises jusque-là segmentées et disjointes : c'est parce que les commerciaux, les juristes, les actuaires, les financiers sont susceptibles d'apparier des informations qu'ils produisaient jusque-là séparément que l'inertie attachée à leur base de donnée pourrait être remise en cause ; mais cette confrontation va sans doute très au-delà des simples difficultés de raboutage entre telle et telle base : si les catégories y sont différentes, c'est aussi parce que les cadres de pensée qui les fondent et les interrogations qui les motivent sont eux aussi hétérogènes et, au moins potentiellement, conflictuels.
- Le dernier enjeu que l'on peut schématiquement évoquer ici est de nature juridique. La question est moins celle de la description du régime juridique de l'usage des données personnelles que celle de ses conditions de production et d'appropriation, en particulier par les entreprises d'assurance. La mise en place du récent « Règlement général sur la protection des données » (RGPD) offre de ce point une situation quasi-expérimentale. L'enjeu sera, d'abord, de reconstituer la séquence qui a présidé à la mise en place de cette nouvelle régulation, et notamment d'y repérer le rôle qu'y ont joué les acteurs économiques les plus susceptibles de faire un nouvel usage de ces données – et donc d'être entravés par les règles qu'on leur impose. De ce point de vue, la littérature oppose classiquement deux schémas causaux qui balisent les extrêmes d'un spectre explicatif qu'il faudra prendre le soin de parcourir : le premier schéma repose sur une logique fonctionnelle (le droit s'énonce pour régler une difficulté que les activités socio-économiques font émerger), le second obéit à une logique politique (le droit vient sanctionner des rapports de force, et les régulations sont l'expression du pouvoir des acteurs les plus puissants : pour une discussion récente et critique de cet argument dit de « capture régulationnelle », voir Carpenter, 2013). Comme l'ont établi de longue date les travaux du courant *Law and society* par ailleurs, un cadre régulationnel ne se résume pas aux principes du droit qui le fondent et aux règles qu'il énonce (Edelman et Suchman, 1997) : il faut s'attacher à décrire comment les acteurs ont (ou n'ont pas) conscience du droit – et à quelles conditions cette conscience se construit, il faut également s'intéresser à la manière dont les entreprises, après que la règle a été énoncée, contribuent à en préciser les déclinaisons, à en lever les ambiguïtés et, éventuellement, à en infléchir les logiques (cf. les travaux de Dobbin (2009) sur le cas emblématique des lois sur les droits civiques).

Le traitement des données

Elucider les conditions de possibilité du déploiement du *big data* n'impose pas seulement de préciser celles qui règlent la collecte et la mise à disposition des données de masse – cela suppose également de repérer les enjeux attachés au *traitement* de ces données. Si l'on retient la caractérisation que nous en proposions plus haut, le propre de ces traitements est de n'obéir à aucun balisage stabilisé *a priori* : il n'existe pas une famille de traitements, dont les contours seraient stabilisés et les propriétés connues, qui auraient vocation à décrire efficacement les données de masse à l'exclusion de tous les autres. Autrement dit, le mouvement actuel autour du *big data* est celui d'une concurrence entre des familles de traitements, souvent issues d'espaces disciplinaires très éloignés, dont la pertinence reste le plus souvent à démontrer et dont nous souhaitons étudier les logiques de circulation : si des traitements concurrents sont mobilisables pour rendre compte des mêmes jeux de données, comment comprendre que certains s'imposent plutôt que d'autres ? L'histoire et la sociologie des sciences et de l'innovation ont depuis longtemps démontré que les solutions qui s'imposent ne sont pas nécessairement les plus efficaces (loin s'en faut ! cf. David, 1985 ; Cusumano et *al.*, 1992) : les phénomènes de dépendance du sentier jouent un tel rôle, en l'occurrence, que certaines options sont privilégiées parce qu'elles sont adoptées de manière plus précoce et que des effets d'irréversibilité les protègent ensuite de toute remise en cause.

Mais s'il y aurait quelques naïvetés à postuler *a priori* que les modèles qui s'imposeront seront les plus efficaces pour décrire les jeux de données massives, il faut s'interroger sur les mécanismes qui joueront en la matière un rôle déterminant dans le monde assurantiel. De ce point de vue, rappelons que, sans être impossibles, les transferts disciplinaires sont rares, qui voient une famille de modèles passer d'un point à l'autre de l'espace scientifique. Le plus souvent, les modèles qui prévalent pour rendre compte de tel ou tel type de données sont marqués au sceau d'une certaine inertie, et s'ils circulent, c'est dans des espaces contigus : dans la reconstitution qu'il propose de la genèse des idées factorielles et de leur diffusion, Olivier Martin montre que le foyer de développement de ces outils se situe au sein de la psychologie, et qu'elles s'étendent à partir de ce cœur vers des disciplines voisines qui contribuent à les désolidariser partiellement des théories qu'elles doivent initialement permettre de prouver (Martin, 1997). On peut certes opposer à ce principe d'inertie des contre-exemples – qui parfois touchent de près les activités assurantielles. Le cas de l'amalgame aléa-hétérogénéité, que nous étudions ailleurs (François et Frezal, 2018), en est sans doute le meilleur exemple. C'est en faisant circuler des raisonnements probabilistes entre l'astronomie (discipline dont il est originaire) vers les statistiques sociales (qu'il entreprend de développer avec une infatigable énergie) que Quételet va le (re)mettre en circulation. Quételet rappelle ainsi qu'un astronome, lorsqu'il effectue une mesure, fait une erreur aléatoire : si on répète l'opération, il est possible d'appréhender la qualité de la mesure en observant la dispersion des résultats et dès lors se rapprocher de la « vraie valeur » en calculant une moyenne. De même, si on effectue de nombreuses mesures du tour de poitrine d'une statue, alors on observera une dispersion de même forme mathématique. Il constate enfin que si l'on mesure le tour de poitrine des soldats d'un régiment, on observe la même forme de dispersion que pour les erreurs de mesures de la statue : le formalisme mathématique est identique pour décrire l'erreur d'estimation d'une grandeur donnée, et pour décrire la dispersion d'une caractéristique au sein d'une population d'individus. La postérité de Quételet fut immense, bien que parfois implicite (Desrosières, 2008a). Elle est surtout loin de se limiter au seul périmètre des statistiques sociales, puisque les innovations conceptuelles avancées par Quételet ont ensuite exercé une influence considérable sur la théorie cinétique des gaz, la « mécanique statistique », les questions relatives à l'hérédité, etc.

On voit donc se dessiner deux formes de circulation possibles, qui balisent les deux pôles d'un spectre : une circulation de proche en proche, nécessairement limitée dans son extension, et des transplantations de loin en loin, qui voient certaines démarches être subverties par des innovations mathématiques mises au point dans des espaces *a priori* très éloignés. L'hypothèse implicite qui semble déterminer les raisonnements, en matière de *big data*, privilégie la seconde et néglige souvent les forces susceptibles d'imposer la première. Nous proposons de constituer cette interrogation en questionnement empirique, et de suivre la manière dont s'organise la concurrence entre les traitements qui entendent décrire les jeux de données massives dans le secteur de l'assurance. Nous privilégierons plus particulièrement l'étude de deux espaces :

- Un *espace de formation*, tout d'abord. C'est un constat trivial : pour que les modèles soient utilisés, il faut que ceux qui les mobilisent y soient d'abord formés. Ce point d'évidence est aussi, potentiellement, un point de (dé)blocage fondamental dans l'histoire de l'usage des outils mathématiques. Si l'axiomatisation proposée par Kolmogorov au début des années 1930 fut en effet au principe d'un usage démultiplié des probabilités, c'est parce qu'elle fut très rapidement transmise et vulgarisée et routinisée dans les manuels de langue anglaise (Van Plato, 1994), et cette vulgarisation des axiomes de Kolmogorov va en particulier jouer un rôle décisif dans le développement des mathématiques financières (Jovanovic, 2012). C'est en effet à cette époque que des économistes, pour certains (comme Markowitz) engagés dans l'étude des marchés financiers en mobilisant les outils de l'économétrie s'appuyèrent sur ce corpus désormais accessible pour donner à leur approche un tour beaucoup plus théorique. Ils mobilisèrent des outils mathématiques auxquels ils pouvaient se former pour modéliser les propriétés des variables aléatoires et pour élaborer les premières propositions de la théorie financière.
  La formation aux questions relatives au *big data* est, aujourd'hui, rien moins que stabilisée. La plupart des institutions d'enseignement supérieurs font du « *big data* », du « digital », des « données et méthodes numériques » une priorité affichée de leur politique d'enseignement et de recherche – sans que très souvent cette priorité ne se décline autour d'axes clairs et immédiatement lisibles. A l'Ecole polytechnique fédérale de Lausanne, la formation au digital est pour l'essentiel localisé dans le département de Sciences humaines – mais elle est entièrement tournée vers les *Computer sciences* et assurée par des spécialistes de mathématiques appliquées, qui n'enseignent qu'une petite partie des outils disponibles pour analyser les jeux de données massives. Si l'on tente de clarifier l'espace des choix qui objectivement se présente aux institutions d'enseignement supérieur, on peut tout d'abord distinguer entre des formations qui privilégient les *méthodes* (comment analyser les données ?) et celles qui tentent au contraire de sensibiliser les étudiants aux *pratiques* (quels nouveaux univers de pratiques s'ouvrent, dans tel ou tel domaine, du fait du développement de ces jeux de données ?). On peut ensuite distinguer, parmi ceux qui décident d'enseigner les méthodes, selon les *familles de méthodes* qu'il s'agit d'enseigner : pour reprendre l'exemple de l'EPFL, que nous avons eu l'occasion d'étudier, sont privilégiées des techniques de *machine learning* très pointues – ainsi que des modèles de simulation de réseau issus des *computer sciences* – tandis que d'autres techniques, plus proches des sciences sociales (analyses multidimensionnelles ou économétrie) sont totalement ignorées. Ces deux espaces de choix (pratiques ou méthodes ? quelles méthodes ?) ne sont que très imparfaitement discernées par les écoles et les universités dont la politique, en la matière, est moins déterminée par des options fermes et réfléchies que par des logiques de recrutement dans un contexte de (très) grande rareté des compétences. Repérer quelque chose comme des logiques structurantes est d'autant plus délicat, à ce stade, que toutes les formations sont concernées : celles qui s'appuient sur les sciences de gestion et les sciences sociales, car le *big data* est censé avoir des conséquences

d'importance sur les pratiques des gestionnaires ou des clients ; et celles dont le cœur de métier est davantage organisé autour des sciences de l'ingénieur (mathématiques, informatique, physique, etc.) car elles y voient l'opportunité de prendre pied, *via* l'outillage technique, sur des territoires qui leur étaient jusque-là moins favorables. C'est donc un monde mouvant qu'il s'agit de décrire, et dont il faut s'attacher à comprendre les transformations.

En matière de formation spécifique aux métiers de l'assurance, nous avons déjà attaché une attention privilégiée au groupe des actuaires – et, sans nous y restreindre, nous prolongerons les interrogations déjà travaillées dans Pilmis (2016) et Ollivier (2017). Les actuaires ont en effet utilisé la réforme Solvency II pour accroître et consolider leur juridiction (Ollivier, 2017) : les enjeux liés à l'appréhension mathématique des risques se sont considérablement accrus et complexifiés avec cette réforme, et les bénéfices qu'en ont retiré les actuaires tant sur leur marché du travail (où les prix ont crû de manière spectaculaire) qu'au sein des organisations (où leur pouvoir s'est affirmé) sont spectaculaires. Le développement du *big data* est vu par nombre d'entre eux – et notamment par ceux qui les forment – comme une remise en cause potentielle de leur position, qui peut aussi se transformer en opportunité. Tels qu'elles peuvent être décrites par Pilmis (2016), les formations dispensées dans les écoles d'actuariat tentent de conserver le socle actuariel – dont les compagnies d'assurance auront toujours besoin – tout en transformant les actuaires en *data scientist*, *i.e.* en élargissant le bagage mathématique auquel ils sont formés à des familles de modèles qu'ils ignoraient jusque-là – en développant les formations à l'algorithmie par exemple – et en renforçant également la formation en programmation, à Python et à R en particulier. Tout l'enjeu pour les actuaires, est celle de savoir s'ils vont parvenir à préserver leur juridiction en en étendant le périmètre – ou s'ils verront au contraire se développer des vis-à-vis disposant de compétences techniques au moins équivalentes aux leurs pour aborder des sujets contigus.

- Un espace de *production*. Comprendre la transformation des espaces de formation est une étape évidemment nécessaire pour tenter de repérer les transformations de moyen et de long terme qui se feront jour au sein des entreprises d'assurance. Mais pour rendre compte des conditions d'appropriation et de circulation des modèles attachés au *big data*, il faut aussi investir un autre espace où des modifications de court terme sont susceptibles de se faire jour : l'espace de la production des compagnies d'assurance, qui au moins pour certaines d'entre elles tentent d'inventer des formes organisationnelles susceptibles de les voir s'approprier de nouvelles pratiques qu'elles peuvent contribuer à inventer. Pour ce que nous avons pu en voir au cours de nos précédentes enquêtes, les dispositifs d'appropriation que déploient les entreprises d'assurance sont encore tâtonnants, incertains et réversibles : on est loin de pouvoir identifier une ou plusieurs formes typiques qui s'imposeraient à l'ensemble d'un champ (Meyer et Rowan, 1977). Autrement dit, de même que Solvency II a pu se saisir comme l'occasion de repérer les modalités d'une innovation organisationnelle, celle de l'invention de la direction des risques (Bizieux et François, 2016), de même l'avènement possible du *big data* dans les entreprises d'assurance doit permettre de repérer les formes organisationnelles que revêt cette innovation technique.

  De ce point de vue, et sans prétendre faire de cette recension sommaire autre chose qu'une esquisse typologique provisoire, nos explorations ont permis d'identifier deux dispositifs dont les destins, quand nous avons provisoirement suspendu nos allers-retours, étaient loin d'être scellés. Le premier, déployé dans un grand groupe d'assurance européen, consiste à constituer un *Lab* organisé autour de deux principes : un relatif

éloignement des tâches de production, d'abord, et une concentration sur des activités de recherche ; le rassemblement en un même lieu (organisationnel, mais aussi physique) de chercheurs issus d'espaces disciplinaires hétérogènes, afin de couvrir l'ensemble des familles de modèles susceptibles d'être mobilisées dans l'analyse des vastes jeux de données. Le second dispositif, lui aussi mis en place dans un grand groupe européen – concurrent du premier – repose sur des principes à peu près inversés : non pas concentrer des chercheurs dans des espaces dédiés à la recherche, mais au contraire les distribuer dans des *task forces* dédiés à des tâches opérationnelles. Reprendre le travail sur les espaces de production pour repérer en leur sein les modalités d'appropriation des enjeux et des outils du *big data* nous permettra de préciser le destin de ces deux dispositifs et de repérer ceux qui depuis se sont inventés ; il nous autorisera aussi à décrire les conditions de leur développement et, éventuellement et suivant des dimensions qu'il faudra préciser, de leur succès – ou au contraire de leur remise en cause et de leur éventuel échec. On s'attachera à montrer comment le destin de ces dispositifs organisationnels est susceptible de créer des effets épistémiques au sein des organisations, *i.e.* de créer des effets de dépendance de sentier quant au développement (ou à l'atrophie) de tel ou tel type de traitement des données de masses.

S'intéresser aux formes d'appropriation qui se font jour au sein de telle ou telle organisation permettra, enfin, de souligner les inégalités susceptibles de se faire jour, sur ce point, entre les entreprises d'assurance. A la différence du secteur bancaire auquel on le compare souvent, le secteur assurantiel est en effet composé d'entreprises d'une très grande hétérogénéité, qu'il s'agisse de leur statut (capitaliste ou mutualiste) ou de leur taille (des groupes immenses, comme Axa ou Allianz, côtoient de très petites structures). Comme ce fût le cas pour l'appropriation de Solvency II, on peut faire l'hypothèse que cette hétérogénéité n'est pas sans incidence sur les modalités d'appropriation du *big data* au sein des entreprises d'assurance : la conscience que les entreprises peuvent avoir des enjeux, les compétences qu'elles peuvent mobiliser (en interne ou sur le marché du travail ou des consultants) sont en effet très inégales. Il conviendra de repérer les conséquences que peuvent engendrer ces inégales capacités d'appropriation sur les formes d'organisation internes des entreprises, mais aussi – et surtout – sur les dynamiques concurrentielles susceptibles de se faire jour au sein du secteur : peut-on imaginer, comme cela semble être le cas pour Solvency II, que ce nouveau choc accélère encore les logiques de concentration qui s'y font jour ? Si ce devait être le cas, on pourrait imaginer que les mécanismes causaux diffèrent en profondeur de ce qui prévalait dans le cas de Solvency II : là où les effets de Solvency II passaient notamment par l'imposition de nouvelles règles relatives à la santé *financière* de telle ou telle organisation, ce serait, par exemple, *commercialement* que les petites structures, incapables de tirer parti des données pour segmenter au mieux leur clientèle, serait cette fois-ci fragilisée.

Un changement de paradigme ?

Après nous être interrogé sur les conditions de possibilité de l'avènement du *big data*, nous souhaitons nous interroger sur ses conséquences, en plaçant au cœur de nos interrogations ce qui semble être l'hypothèse indiscutée de tous les discours qui y sont consacrés : le *big data* change-t-il quelque chose, et si oui que change-t-il et à quoi ? En formulant la question centrale de notre démarche de manière délibérément triviale, nous entendons avant tout en souligner le caractère fondamental : le *big data* est-il effectivement au principe d'un changement de paradigme dans nos pratiques – ou les changements qui lui sont attachés demeurent-ils encore hypothétiques, voire improbables ? Rappelons ce qui motive l'hypothèse d'un changement de paradigme : le recours aux données massives permettrait d'affiner considérablement la prédictibilité des comportements individuels, et par conséquent des risques encourus par chaque individu. Cette hypothèse, qui doit elle-même être discutée, doit notamment être évoquée quant à ses implications : dans quelle mesure l'avènement du *big data* est-il susceptible d'avoir une incidence sur les opportunités et les contraintes des parties prenantes du monde de l'assurance ? Nous discuterons successivement trois niveaux : l'échelle individuelle de l'assuré, celle, organisationnelle, de l'entreprise d'assurance, de sa politique de tarification et de son *business model*, et celle, sociale, de l'Etat-providence et de la solidarité qu'il engage.

L'assuré et l'échelle individuelle

Angle théorique : des modèles gaussiens aux modèles quantiques

Traditionnellement, l'usage des probabilités en assurance s'appuie sur une approche fréquentiste : la probabilité de sinistre est calculée sur un grand nombre d'assurés, supposés présenter des caractéristiques similaires et un risque homogène. Cette approche dite gaussienne suppose une distribution normale des variables de l'analyse et met en jeu essentiellement des moyennes. La loi des grands nombres joue sur l'ensemble de la population (ou du segment), pour assurer la réalisation en moyenne de ces variables aléatoires, comme espérance de sinistre par exemple. Le résultat du modèle n'a pas de sens au niveau individuel et n'est considéré performant qu'en fonction de sa capacité à modéliser les risques de la population dans son ensemble.

L'usage des probabilités inhérent aux techniques du *big data* est de nature différente : il ne s'agit plus de prévisions macro-économiques des résultats d'une population, mais de la prédiction d'un comportement individuel. Comme décrit dans la partie précédente, la multiplication des données accessibles sur chaque individu rend en effet possible une nouvelle forme de calcul, qui ne suppose aucune distribution particulière ni ne cherche à dégager une formule valable sur l'ensemble : la probabilité calculée par le modèle, ou score de l'individu, se veut être un indicateur de probabilité d'occurrence de l'événement pour cet individu spécifique (Breiman 2001 ; Siegel 2016). Même s'il convient de distinguer entre les différents types de modèles (un score obtenu par arbre de décision n'étant pas exactement de même nature qu'une classification par réseau neuronal), on peut cependant avancer que l'interprétation de cette probabilité individuelle est beaucoup plus problématique que celle obtenue sur la population dans son ensemble ; alors que cette dernière s'inscrit dans l'approche fréquentiste décrite plus haut, le score individuel, de par la perspective purement individuelle qu'elle adopte, contredit cette approche.

Ce basculement d'une forme de probabilité à une autre semble renforcer l'amalgame entre des phénomènes intrinsèquement aléatoires (tels que la survenance d'un sinistre pour un individu spécifique), et des phénomènes pluriels qui acquièrent une régularité sur le grand nombre et dont seule la mesure est aléatoire (telle que la fréquence de sinistres observée sur une population). On retrouve ici, dans un nouveau contexte, l'amalgame entre aléa et hétérogénéité qui a déjà fait l'objet de travaux de la chaire (Frezal 2015 ; Frezal 2018). Ces travaux pourront être mis à profit

et complétés par une analyse des modèles spécifiques au *big data* dans la perspective de cette distinction entre aléa et pluralité.

Sera examinée notamment la pertinence d'un passage d'une statistique gaussienne à une statistique dite quantique : les techniques prédictives conduisent en effet à une individualisation de la probabilité, ou du moins à l'attribution d'un score dont il convient de questionner le sens. Une analogie semble formellement possible avec la physique quantique, dans laquelle des probabilités (de présence) sont attribuées à des particules et les prédictions sont aussi, comme dans le cas des modèles de data science, de nature probabiliste. Les débats autour du sens à donner à ces probabilités quantiques montrent cependant la difficulté conceptuelle de cette approche (Boyer-Kassem 2015). Sur le plan humain, ces prédictions sont particulièrement problématiques car elles tendent, paradoxalement, à réduire le comportement à un indice pris pour une indication déterministe (Harcourt 2007). Une étude sociologique de l'usage des scores permettra d'orienter cette recherche.

Angle sociologique

Pour évoquer les hypothèses qui fondent notre travail sur l'assuré et sur les évolutions attachées au *big data* qui sont susceptibles de l'affecter, nous partirons du constat que si l'attribution d'un risque individualisé à tel ou tel agent économique est peut-être une nouveauté pour l'assurance, cette pratique existe parfois de très longue date dans d'autres secteurs économiques, et notamment dans celui des activités de crédit. La littérature sur ces questions est abondante (pour une synthèse, voir Rona-Tas et Guseva, 2018) et c'est en partant de ses résultats que nous souhaitons construire nos interrogations.

La pratique du *credit rating* ou du *scoring* consiste, explique Carruthers (2013), à tenter de transformer les situations d'incertitude en situation simplement risquée. Selon les distinctions canoniques de Knight (1985), un individu peut être amené à prendre des décisions dans trois situations bien distinctes : il peut faire face à des situations de certitude (il sait ce qui va se passer, il peut donc décider en connaissance de cause), de risque (il ne sait pas ce qui va se produire, mais il connaît la liste des scénarios possibles et peut leur affecter un coefficient de probabilité : l'avenir n'est pas garanti mais il demeure calculable), ou enfin d'incertitude : dans ce dernier cas, soit les probabilités affectées aux différents scénarios sont inconnues, soit la liste même des scénarios est inconnue – « nous ne savons pas, tout simplement », écrit Keynes (1937) pour décrire ces situations. Dans ces situations d'incertitude radicale, le calcul est impossible, l'exercice de la rationalité devient extrêmement délicat. Pour faire face à ces situations, estime Keynes, les acteurs économiques s'appuient sur des « conventions », *i.e.* des schémas de comportements stabilisés et partagés, qui guident l'individu dans son action quand il ne peut pas calculer. Carruthers (2013) souligne l'importance d'une autre technique : l'acteur économique peut tenter de transformer l'incertitude en risque – il peut, autrement dit, s'efforcer de rendre la situation calculable, par exemple en affectant aux différents scénarios des coefficients de probabilité. La difficulté peut cependant surgir lorsque ceux qui prennent la décision oublient le mouvement qui fonde leur calcul : « *without due acknowledgement of the shift from underlying uncertainty to ostensible risk, decision-makers may be overly confident in their ability to make the right choice. They will also be thoroughly blindsided if the assumptions they made to render uncertainty tractable prove false or insufficiently robust.* » (Carruthers, 2013, p. 526).

Cette caractérisation théorique du *scoring* et des risques qu'il engage peut être précisée, d'abord en revenant sur les opérations de catégorisation que supposent les pratiques de *credit rating*. Rona-Tas et Guseva (2018) expliquent ainsi que le « « *scoring takes individuals out of their social context (…) and disaggregate them into a finite number of variables, making them "dividuals" (…). These characteristics can include a wide variety of socioeconomic, demographic, and other attributes or only those related to credit histories. Then, through statistical manipulation, the person is connected by a series of comparisons to a large group of*

strangers equally disaggregated into a handful of variables. Finally, he or she is reassembled into a single number or category. » (Rona-Tas et Guseva, 2018, p. 9). Les scores attribués aux différents emprunteurs masquent, dans leur évidente simplicité, l'ensemble des choix qui président à leur attribution :

- Ces choix portent, d'abord, sur *la nature des variables* qui peuvent être engagés dans la définition des scores. Aux Etas-Unis, seule l'histoire du crédit de l'individu est prise en compte (a-t-il été régulier dans ses remboursements ? a-t-il fait défaut ?), tandis qu'en France et au Royaume-Uni, une série de variables sociodémographiques peuvent être prises en compte (le revenu, l'âge, le statut marital), dont la plupart sont considérés comme des indicateurs de la stabilité des emprunteurs. En Chine, les données utilisées pour attribuer un score aux individus viennent d'un spectre d'expériences plus large encore, puisqu'y sont inclus le fait de tenir sa parole et de se conformer aux règles légales et morales ainsi qu'aux standards professionnels – l'enjeu des scores est alors de mesurer la fiabilité de l'individu.
- Ces choix portent également sur *l'algorithme utilisé* pour traiter ces données – algorithme qui est le plus souvent opaque : les prêteurs le cachent à leurs concurrents dans l'espoir de disposer d'un modèle qui leur permettra de mieux prédire le comportement de leurs clients potentiels ; mais ils le cachent aussi à leurs clients, afin de les empêcher de jouer avec le système (Rona-Tas et Hiss, 2010). Cette opacité de l'algorithme est loin d'aller de soi – elle fait d'ailleurs l'objet de contestation juridique, au Brésil ou en Allemagne notamment : dans les situations où, comme au Brésil, l'ensemble des variables prises en compte dans l'attribution d'une note doivent être communiquées au client, comment justifier que l'algorithme qui en assure le traitement lui soit masqué ? La décision de justice, qui préserve l'opacité de l'algorithme, est fondée sur le fait que le modèle n'est pas une donnée mais une équation mathématique : il n'ajoute aucune information aux données et se contente de les réagencer (Doneda, 2016).
- Un dernier espace de choix renvoie à *l'usage* qui peut être fait des scores : comme nous le rappelons ailleurs, un outil comme le *scoring* ne vaut que pour l'usage qui en est fait (François, 2011). En France, par exemple, les emprunteurs potentiels peuvent être acceptés ou rejetés (Lazarus, 2012), tandis qu'aux Etats-Unis le score du candidat au prêt ne le disqualifie pas nécessairement : il peut emprunter mais paie plus cher. C'est sans doute en Chine que les scores attribués aux individus sont utilisés le plus largement (Chen et Cheung, 2017) : un mauvais score ne peut exclure quiconque d'un prêt, mais il peut lui interdire de prendre l'avion ou le train à grande vitesse, ou d'accéder à un emploi de fonctionnaire. Ces usages ne sont par ailleurs pas figés : aux Etats-Unis, les scores interviennent de plus en plus fréquemment lorsqu'il s'agit de juger de la qualité d'un individu qui souhaite assurer sa voiture, louer un appartement ou se faire embaucher (Fourcade et Healy, 2017).

Pour être correctement décrites et caractérisées, les pratiques de *scoring* qui pourraient se faire jour dans les entreprises d'assurance doivent détailler les choix qui sont effectués sur ces trois dimensions – comme doivent être décrits les processus qui mènent à privilégier telle ou telle option. Le simple balisage de ces espaces de choix permet par ailleurs de rappeler tout ce qu'ils peuvent engager de *politique* : le recours au *scoring* n'est pas une nécessité ou une fatalité technologique, il est le résultat d'une prise de position explicite qui, le plus souvent, se donne à voir dans les lois qui l'autorisent et qui en définissent les termes. Les travaux sur le *credit rating* insistent ainsi beaucoup sur la détermination politique et juridique des pratiques de *scoring*. Ainsi, par exemple, de l'origine même du dispositif qui, aux Etats-Unis, fut mis en place pour lutter contre les discriminations dans l'accès au crédit dont étaient victimes les minorités de couleur : les notes attribuées aux emprunteurs potentiels devaient permettre de s'opposer aux décisions

discrétionnaires de certains prêteurs (Pager et Shepherd, 2008). Ainsi, également, des limites (théoriques) attachées à l'automaticité des décisions, que l'Union Européenne interdit en imposant que les décisions de crédit fassent toujours intervenir un humain – même si cette règle est contournée dans la plupart des pays (Rona-Tas et Guseva, 2013). L'analyse des pratiques en matière de *scoring* et l'anticipation de leur extension dans le secteur de l'assurance doit impérativement faire intervenir cette dimension politique et s'attacher à décrire l'ensemble des acteurs qui interviennent dans le processus qui étend ou restreint le spectre du recours légitime à ces dispositifs d'individualisation de la mesure du risque.

Si les travaux sur le *credit rating* permettent de battre en brèche l'hypothèse d'une inéluctabilité technologique de ces pratiques, ils relativisent aussi l'idée d'une automaticité des décisions : le crédit demeure une activité relationnelle, ce qui apparaît clairement quand la relation se déploie dans de petites communautés (Holmes et al. 2007) – mais c'est aussi vrai dans des grandes banques (Li et *al.*, 2009). Les relations de face-à-face sont particulièrement importantes aux deux extrêmes de la hiérarchie sociale : les banques les préservent pour leurs clients les plus cossus (Rona-Tas et Guseva, 2014), mais on les retrouve aussi à l'autre extrême du spectre social dans les entreprises de crédit immobilier ou de micro-finance (Stenning et *al.* 2010). Dans quelle mesure les mécanismes qui fondent l'importance des relations de face à face sont-elles susceptibles de se retrouver pour les prestations assurantielles ? L'enquête devra s'attacher à le déterminer. Mais les résultats sur la relation de crédit alertent contre le risque de prêter au dispositif une efficace qui ne dépend, à la fin du compte, que de l'usage de ceux qui s'y réfèrent.

L'entreprise d'assurance, ou l'échelle organisationnelle

Angle sociologique et historique

L'une des principales interrogations soulevées par l'avènement du *big data* porte sur l'éventuelle remise en cause du modèle économique des compagnies d'assurance occidentales. Selon certaines hypothèses, cette remise en cause pourrait être équivalente à celles que connurent ces mêmes compagnies au cours du XVIII^{e} siècle lorsque, pour prendre la mesure des risques qu'elles prenaient et des rendements qu'elles pouvaient en espérer, les compagnies d'assurance ont cessé de s'en remettre à une appréciation intuitive et qualitative pour s'appuyer sur le calcul des probabilités et, notamment, sur la loi des grands nombres. Au XVIII^{ème} siècle, en effet, l'assurance était considérée comme un jeu et les entreprises d'assurance disqualifiées comme des entreprises de pari (Daston, 1988, p. 164-165), tant du point de vue des souscripteurs – qui achetaient moins une protection qu'ils ne jouaient à la loterie lorsqu'ils souscrivaient une assurance vie par exemple (ou bien lorsqu'ils s'assuraient contre l'adultère ou le mensonge) – que du point de vue des compagnies elles-mêmes, qui ne s'appuyaient pas sur des statistiques pour tarifer leurs contrats. Au mieux, les compagnies les plus sérieuses s'appuyaient, à l'instar des compagnies italiennes d'assurance maritime du XV^{ème} siècle, sur des souscripteurs dont l'expérience et la connaissance du contexte permettaient de porter une appréciation qualitative sur la qualité du risque (l'expérience du capitaine, le niveau de piraterie en cours, la dangerosité du trajet, etc.), ou bien, comme les fournisseurs de rente hollandais du XVIème siècle, sur une visite médicale (individuelle) plutôt que sur une table de mortalité (collective). A l'opposé de cette dimension aléatoire, la théorie mathématique en cours de développement offre, avec la loi des grands nombres formulée par Bernouilli, un monde « simple, stable *et prévisible* » (Daston, 1988, p. 113). L'adoption progressive de ces outils renverse la conception du risque : on considérait auparavant que plus il y avait d'assurés ou plus l'horizon temporel était éloigné, plus il y avait d'incertitudes ; la conviction s'impose désormais que plus les assurés sont nombreux ou plus l'horizon temporel est profond, mieux s'applique la loi des grands nombres, et plus les compensations permettent de réduire l'incertitude. Ce changement dans l'appréhension du risque permet d'ouvrir très largement le marché de l'assurance : si plus il y a d'assurés, mieux le risque

est contrôlé, alors les entreprises d'assurance ont intérêt à assurer un très grand nombre d'acteurs plutôt qu'à réserver leurs services à quelques *happy few*.

Cette ouverture de l'offre suppose, cependant, de trouver face à elle une demande – ce qui est loin d'aller de soi. Lorsque les entreprises d'assurance étaient assimilées à des bureaux de pari, elles n'étaient en effet pas les seules à être disqualifiées – leurs usagers l'étaient aussi : ils étaient des joueurs qui, lorsqu'ils s'assuraient sur la vie, pariaient sur l'hypothèse de leur propre mort. Dans son livre désormais classique, Viviana Zelizer montre comment l'ouverture véritable du marché de l'assurance – et la mise en place, au sein des compagnies, d'un nouveau *business model* fondé sur le nombre plus que sur l'hypersélection des assurés – a supposé que soit réalisé un très important travail normatif, notamment de la part des pasteurs de la côte Est (Zelizer, 1979). Même si sa chronologie n'est pas toujours parfaitement explicite, deux temps s'y opposent. Dans un premier temps, s'assurer constitue un péché d'une extrême gravité : vouloir se prémunir contre les risques que peuvent courir ses proches si l'on en vient à mourir, c'est tenter de contredire les desseins de Dieu. Dans un pays, les Etats-Unis, où le taux de mortalité (masculine en particulier) est très élevé pour qui passe la frontière des Appalaches, et où les solidarités traditionnelles ont été à peu près anéanties par les trajectoires migratoires, les risques encourus par les survivants sont cependant immenses et des trajectoires biographiques dramatiques vont amener les pasteurs à entièrement réviser leur construction normative : s'assurer, ce n'est plus parier contre Dieu, c'est au contraire prendre soin de ses proches – et il faut encourager cette pratique qui, dès lors, n'a plus rien d'un jeu de hasard.

Ce déplacement, lent à se mettre en place – les évolutions pointées par Zelizer sont perceptibles à la fin du XVIII^{e} siècle au Royaume Uni, elles s'imposent aux Etats-Unis dans le dernier XIX^{e} siècle – ouvre sur une période de fonctionnement stabilisé des compagnies d'assurance, dont nous souhaiterions étudier en détail la politique de tarifications. Les principes théoriques en sont connus : appuyé sur les outils probabilistes et la formalisation qu'a pu en donner l'économie de l'assurance (inversion du cycle, mutualisation, différence vie et non vie, etc.), le modèle économique des entreprises d'assurance est connu et stabilisé depuis des décennies. Ses déclinaisons empiriques le sont moins : comment cette économie très particulière se décline-t-elle dans les pratiques quotidiennes des acteurs qui travaillent au sein des entreprises, dans les dispositifs organisationnels qui structurent leurs pratiques, dans les chaînes de décision et de contrôle où ils s'inscrivent, dans les contrats qu'ils mettent au point et qu'ils proposent aux assurés ? Cette question a été nettement moins travaillée. On peut y voir, sans nul doute, l'un des symptômes du déficit d'attention dont le secteur assurantiel a fait l'objet – en particulier si on le rapproche du secteur bancaire ou des marchés financiers. L'étude empirique des entreprises assurantielles, si elle s'est récemment développée (voir par exemple Chan, 2009 ; Jarzablowski et *al.*, 2015 ; Yates, 2008), demeure à bien des égards le parent pauvre des travaux de sciences sociales consacrés au secteur financier – ce qui ne laisse pas de surprendre si on se rappelle du poids économique du secteur assurantiel. Nous proposons de soulever ces questions fondamentales d'appréhension empirique du *business model* des entreprises d'assurance traditionnelles en faisant des pratiques de tarification une porte d'entrée – un traceur, si l'on veut – permettant de décrire l'ensemble de ces pratiques et de la division du travail où elles s'inscrivent, des rapports de pouvoir et des conflits qu'elles engendrent, de leurs modes de résolution et des redéploiements qu'ils provoquent. Nous contribuerons ainsi à une analyse empirique des modes de fixation des prix qui continue d'être le parent pauvre de la sociologie des organisations et des marchés – en dépit de travaux pionniers (Baker, 1984) ou plus récents (Finez, 2014) (pour une revue, voir Filleule, 2008 ; Beckert, 2011). L'étude des pratiques de tarification, appuyées sur des fonds d'archives et des campagnes d'entretiens menées avec de grands témoins du secteur permettra de reconstituer les chaînes d'interaction au cœur du fonctionnement des entreprises d'assurance, dans une configuration définie par la massification des produits

d'assurances et l'appui sur le principe de mutualisation. On pourra ainsi contraster plus précisément ce mode de fonctionnement – à ce jour, répétons-le, encore très insuffisamment documenté dans ses déclinaisons concrètes – avec celui qui pourrait être engendré par la généralisation du recours aux données massives.

Le chercheur en sciences sociales s'aventure ici sur un terrain qui lui est peu familier : les transformations dont il est ici question sont encore très largement à venir, et il s'agit bien plus d'accompagner, en les observant pour en prendre une juste mesure, des mutations en cours que de les documenter rétrospectivement comme le sociologue ou l'historien sont plus classiquement habitués à le faire. Si donc l'on tente d'esquisser les directions plus ou moins fréquemment évoquées dans lesquelles pourrait se redéployer le modèle économique séculaire des compagnies d'assurance, on peut en évoquer deux principales.

- La première, la plus classique peut-être, nourrit une récente livraison de la revue *Risques* (2015, n°103). Elle repose sur l'hypothèse que la mobilisation des données massives par les assureurs est susceptible de remettre en cause la devise que la Lloyds s'était donnée en 1688, et qu'auraient pu depuis faire leur les assureurs des pays occidentaux : « *The contribution of the many to the misfortune of the few* ». L'économie classique de l'assurance est confrontée, explique Lasry (2015), à un dilemme que le *big data* va exacerber. D'un côté, mieux connaître le risque permet de mieux le prévenir et de mieux le tarifer : c'est une logique de segmentaion. D'un autre côté, l'assurance repose sur un principe de mutualisation (tous les assurés payent, et le risque de quelques-uns se réalisent) qui suppose, pour se maintenir, qu'on ignore qui précisément sera soumis au risque – il repose, autrement dit, sur une forme d'ignorance. Le *big data,* explique Lasry, radicalise la tension entre les deux termes du dilemme : on peut imaginer que l'ignorance est susceptible de se réduire au point de disparaître, et qu'il sera bientôt possible de segmenter de plus en plus finement les tarifs pour les faire correspondre aux risques courus par tel ou tel individu.

  L'assurance auto avec boitier Usage Based Insurance (UBI) pourra ici servir d'exemple, bien que de plus en plus d'applications sur smartphone permettent d'obtenir des résultats similaires dans de nombreux domaines (santé, vie…). Ces boitiers collectent en temps réel des données sur la vitesse, l'accélération en trois dimensions et la localisation du véhicule ; ce sont des milliers de mesures pour chaque trajet effectué, là où auparavant n'était accessible qu'un nombre limité de paramètres, obtenus au moment de la souscription du contrat. Ces données permettent de calculer une note de conduite sur la base d'une probabilité de sinistre "personnalisée", résultant de ces seules données individuelles. Dans un cas limite, on pourrait envisager une tarification sur la base de ces seules données télématiques, en contournant totalement la segmentation traditionnelle. On pourrait même considérer une tarification ajustée en temps réel en fonction de l'évolution de la conduite de l'assuré. Dans ces conditions, autrement dit, la segmentation prendrait entièrement le pas sur la mutualisation (voir également Hay, 2015 ; Thourot et *al.*, 2015 ; Charpentier et *al.*, 2015) et bouleverserait ainsi l'équilibre qui a construit le *business model* des entreprises d'assurance.

- Une autre source de déstabilisation vient de la remise en cause ce que l'on pourrait nommer le contrat moral qui s'établit entre l'entreprise d'assurance et son assuré. Cette relation assureur-assuré est, classiquement, une relation asymétrique qui profite à l'assuré : il connaît mieux son risque que l'entreprise à laquelle il s'adresse (on est en situation

d'asymétrie d'information), et il est par ailleurs susceptible d'adopter des comportements que l'entreprise ne peut connaître et contrôler, qui sont eux aussi susceptibles d'accroître le risque qu'il porte. Ce contrat moral et l'asymétrie qui lui est propre sont connus des entreprises d'assurance, qui les acceptent car elles savent qu'elles peuvent en supporter le coût. Or, la mobilisation de données massives par les compagnies d'assurance sont susceptibles de très profondément redéfinir les termes de ce contrat moral. On peut en effet imaginer que l'assureur, qui disposera(it) de très nombreuses informations sur son client et de la capacité d'en tirer une compréhension qui lui échappe, en sache désormais plus sur l'assuré que l'assuré lui-même : l'asymétrie d'information serait alors inverse. La méfiance (la « sélection adverse », dans le vocabulaire des économistes de l'information) qui traditionnellement prévaut surtout entre l'assureur et l'assuré qui peut lui cacher des éléments compromettants s'inverse également : ce serait désormais l'assuré qui se méfierait de son assureur, susceptible de lui cacher des bonnes nouvelles issues de sa collecte de données. L'aléa moral, enfin, pourrait se réduire considérablement – puisque la capacité de surveillance que l'on prête aux outils connectés est potentiellement infinie, l'assuré peut beaucoup plus difficilement masquer à son assureur des éléments de son comportement. Le contrat moral qui a longtemps régi la relation entre l'assureur et l'assuré serait ainsi profondément remis en cause.

On le comprend : le déplacement (la disparition ?) de l'équilibre segmentation-mutualisation d'une part, le redéploiement (l'inversion ?) du contrat moral assureur-assuré d'autre part, sont susceptibles de bouleverser en profondeur le modèle économique séculaire des entreprises d'assurance. Si ces évolutions devaient être avérées (rappelons que l'on n'en est, à ce stade, qu'à des projections spéculatives), deux scénarios d'évolution sont repérables dans la littérature. Le premier consiste en une stratégie de restauration ou de préservation. Il repose sur l'hypothèse selon laquelle l'ignorance, qui limite les possibilités de segmentation et qui continue de rendre possible la mutualisation, peut être délibérément entretenue – en particulier par un régime d'interdictions juridiques : si l'on interdit de discriminer les individus en vertu de leur sexe, de leurs mœurs, de leur handicap, de leur état de santé, de leurs opinions politiques, de leurs caractéristiques génétiques, du lieu de résidence, de leur profession, de leur niveau de richesse – alors le potentiel de segmentation et les risques de dé-mutualisation seront considérablement amoindris. Le second scénario consiste non plus à restaurer ou à préserver, mais à redéployer, éventuellement en relativisant : redéployer, en définissant de nouvelles formes de tarifications qui tirent profit du *big data* sans entièrement faire disparaître les logiques de mutualisation (cf. Thourot et *al.*, 2015). Cet équilibre est d'autant plus susceptible d'être atteint que des travaux soulignent que le recours aux données massives apporte son lot de difficultés (Zajdenweber (2015) montre par exemple que de petits portefeuilles composés d'assurés identiques ont une volatilité accrue si on les compare à ceux plus gros composés d'assurés plus hétérogènes), tandis que des difficultés traditionnelles auxquelles sont confrontées les assureurs ne disparaîtront pas avec le *big data* (Charpentier et *al.*, 2015).

- La deuxième ligne d'évolution possible est moins fréquemment évoquée dans la littérature. Elle mérite à notre sens d'être évoquée pour ce qu'elle s'appuie sur une hypothèse alternative à celle que nous venons d'évoquer, qui faisait reposer ses raisonnements sur le redéploiement de la politique de tarification et du lien assureur-assuré. Selon Berbain et Salamanca (2015), cette prémisse est fragile, car elle suppose résolue une difficulté immense – dont nous faisions l'écho plus haut – liée à la gestion des

données par les assureurs. Les données clients sont en effet collectées par une multitude d'interfaces (réseaux salariés, courtiers, agents indépendants, téléphone, internet, etc.) dont les intérêts ne convergent pas toujours avec ceux de l'assureur qui a recours à leur service. Seules sont donc transmises les informations nécessaires au maintien de la relation entre l'assureur et son apporteur d'affaires : les données clients récupérées par les assureurs sont hétérogènes, souvent incomplètes, parfois erronées. A ces difficultés liées à l'organisation de la collecte s'ajoute l'âge des données : les données dont disposent les assureurs sont souvent très anciennes, différentes nomenclatures y cohabitent et s'empilent dans les systèmes d'informations, rendant l'appariement des bases délicates, voire impossibles. Les modalités de constitution des grands groupes d'assurance accroissent encore les difficultés : parce qu'ils se sont très souvent bâtis par croissance externe, s'y accumulent des systèmes d'informations différents, fragmentés et malaisément combinables. Enfin, Berbain et Salamanca soulignent que la richesse des données dont disposent les compagnies d'assurance ne doit pas être surestimée : une quarantaine de variables sont utilisées dans les contrats d'assurance auto, une vingtaine en MRH et moins de dix en santé. Sans doute disposent-elles, en interne, de données supplémentaires – celles, par exemple, qui portent sur la sinistralité – mais elles sont à ce jour encore peu exploitées. Quant aux données externes, si certaines – comme les tables de mortalité – ont été au principe de la mise en place du *business model* des entreprises d'assurance, elles demeurent encore peu exploitées et leur exploitation systématique risque d'être entravée à l'avenir par une série de proscriptions juridiques.

L'hypothèse selon laquelle la politique de tarification et le lien assureur-assuré vont être profondément remodelées par la mobilisation des données massives est donc fragile. Elle néglige par ailleurs une autre modification, qui pour Berbain et Salamanca est beaucoup plus vraisemblable – et beaucoup plus fondamentale. Le développement du numérique s'accompagne en effet d'une transformation fondamentale des modes de consommation : dans l'exemple que prennent Berbain et Salamanca, la consommation de services automobiles passera moins par l'acquisition d'un véhicule que par la location, de loin en loin, d'une voiture que l'on ne possèdera plus. Ainsi, notent-ils, « de nouveaux acteurs de dimension mondiale pourront au travers de systèmes informatisés proposer une gamme de services ciblés et accéder aux utilisateurs à large échelle. (…) L'avenir de l'assurance automobile est donc fortement remis en cause par un ensemble de facteurs qui viennent réduire la masse assurable ainsi que les marges qui peuvent en être retirées. » (p. 33). D'une manière plus générale, avancent-ils, les biens et services placés au cœur des métiers traditionnels de l'assurance sont susceptibles d'être entièrement recomposées par le développement du numérique et les transformations des modes de consommation qui l'accompagnent. Ces mutations provoquent un déplacement des assureurs dans les chaînes de valeur et contribuent à réduire leur marge : pour poursuivre sur l'exemple automobile, l'assureur ne sera plus en contact direct avec l'assuré, mais avec l'entreprise qui lui loue ponctuellement une voiture et qui fournit, avec le véhicule, une assurance qui le couvre pendant qu'il l'utilise. Sans doute le loueur se retournera-t-il vers un assureur pour mettre en place cette assurance – mais le rapport de force entre les deux parties sera sans doute moins favorable à l'assureur qu'il ne l'était lorsqu'il vendait ses contrats à une multitude de clients. Si l'on suit leur raisonnement, l'assureur recule dans la chaîne de valeurs et ses marges se réduisent.

Angle pragmatique

Nonobstant ces difficultés, certains modèles ou applications pour smartphones commencent à se faire jour dans divers secteurs de l'assurance qui font un usage de données de masse. Nous proposons d'étudier ces modèles, dans la mesure des informations disponibles; modèles en assurance automobile, MRH, santé, vie ou encore catastrophes naturelles.

Le but de ces études sera, de façon provisoire et non exhaustive, de mieux cerner :

- Quel usage est fait des données de masse ? Dans le cas des modèles de santé et des modèles automobile, sont mis à disposition des assurés des applications qui gèrent l'information, synthétisent les données en quelques indicateurs clefs et, dans certains cas au moins, donnent des conseils pour améliorer le score obtenu. Du côté des utilisateurs, sont mis en avant les avantages d'une compréhension quantifiée du soi (au cœur du mouvement « Quantified Self », voir notamment Lupton (2014, 2016)) ; du côté de l'assureur, les bénéfices semblent venir d'une sinistralité plus faible (Patel et al. 2010) mais surtout d'une amélioration des taux de résiliation, grâce à l'engagement accru de l'assuré. Or cet engagement est le plus souvent obtenu par la création et l'alimentation d'un phénomène addictif (Eyal 2014). On peut alors se demander quels sont les principes aux fondements de ces indicateurs ; quels types de comportements encouragent-ils? En incitant à modifier une conduite, ces applications visent à diminuer les risques ; mais le modèle économique sous-jacent s'appuie-t-il réellement sur leur mitigation?
- A l'inverse, les modèles prédictifs font un usage accru de théories issues de l'économie behavioriste ; ces dernières visent à mieux décrire le comportement des agents, qui ne sont plus supposés rationnels. Par là, les nouveaux modèles se donnent-ils les moyens d'influer sur les comportements ? Pourrait-on ainsi envisager de distinguer, au cœur des risques, la part comportementale de l'aléa pur de survenance ?
- Certaines données, notamment les données géographiques, résistent mieux que d'autres à l'individualisation. Une réflexion sur la prise en charge des catastrophes naturelles pourrait permettrait de mieux cerner les principes d'une mutualisation intelligente, à partir des données prises en compte dans le modèle.

De façon plus générale, ces études visent à mettre en évidence les limites pratiques des modèles prédictifs et leur usage à bon escient : tarification, aide à la prévention et/ou service ? Nous le disions plus haut : le terrain, ici, est meuble – les pratiques sont encore mal établies et peu stabilisées, et les nouvelles formes économiques qui se feront jour sont encore sinon à inventer, du moins à consolider et à entièrement déployer. L'enjeu des travaux empiriques que développera la Chaire sera moins de tenter de trancher *a priori* entre tel ou tel scénario – les sciences sociales ne sont pas équipées pour ce type d'exercice – que d'assister à leur déploiement et d'en documenter précisément la compréhension. Au moins voit-on ici quelques lignes de force se dégager à titre d'hypothèses, qui pourront, dans un premier temps au moins, contribuer à guider le regard.

L'Etat providence, ou l'échelle nationale

Les transformations sociétales induites par le *big data* sont parfois qualifiées de révolution (Mayer-Schönberger & Cukier 2013; Brynjolfsson & McAfee 2016), voire de quatrième révolution industrielle, les précédentes ayant été portées respectivement (et rudimentairement) par l'invention de la machine à vapeur à la fin du XVIIIe siècle, l'électricité à la fin du XIXe et l'électronique de la fin du XXe . Même si l'on doute de la pertinence de la comparaison, elle permet de rappeler que la révolution industrielle au sens large a été rendue possible par la mise en place des Etats providence, qui ne venaient pas tant apporter les correctifs nécessaires aux effets néfastes de l'industrialisation qu'offrir un espace favorable à son développement (Ewald 1986,

373). Ces mécanismes, institutionnalisés dans la plupart des pays occidentaux après la seconde guerre mondiale, émergent pour Ewald d'une nouvelle philosophie du contrat social, une « doctrine de la solidarité » déjà formulée à la fin du XIXe siècle, notamment par Léon Bourgeois. On peut distinguer avec Ewald cinq piliers à cette doctrine, chacun ayant subi au cours de son histoire des ébranlements divers :

- Au fondement de cette doctrine se trouve tout d'abord une caractérisation du lien social dans les sociétés modernes comme interdépendance des individus. Cette solidarité « organique » se distingue pour Durkheim de la solidarité mécanique des sociétés traditionnelles qui maintenaient ensemble des communautés relativement homogènes et localisées (Durkheim 1967). La société moderne est une constituée d'individus dissemblables mais qui, de par la division du travail social, se trouvent impliqués les uns par rapport aux autres telles les composantes d'un organe. La métaphore de la maladie contagieuse, se propageant par les microbes que Pasteur découvre à la même époque, illustre l'aspect systémique et complexe des ramifications qui unissent les individus les uns aux autres. Pour être efficace, le combat ne peut être que collectif (Ewald 1986).

- Une nouvelle conception de la responsabilité en découle : les « maux sociaux » mis en avant par Léon Bourgeois dans sa théorie du solidarisme ne relèvent pas de la faute individuelle ; ils se propagent comme les microbes et sont, dans les termes de Durkheim, des « faits sociaux », qui n'ont de sens que collectif, à cet autre niveau de réalité que constitue la population (Foucault 2004). Pour le libéralisme classique, l'accident relevait de la responsabilité individuelle et devait être géré par les individus. La prévoyance était affaire de morale et relevait de la responsabilité de chacun (Laurent 2018). Le basculement d'une conception de l'accident comme sort individuel à celle d'un « fait social » bouleverse cette notion de responsabilité et met au cœur de la régulation de ces événements les mécanismes d'assurance.

- Ces derniers consacrent la lente mise en place d'un mode spécifique d'appréhension de l'aléa ; dans les termes de Knight, l'incertain (non mesurable) laisse la place au risque, modélisable et mesurable (Knight 1985). Mais le risque mesurable et mesuré est collectif : cette appréhension moderne de l'incertitude est ainsi fondée sur le calcul statistique, inséparable composante de la constructions des Etats occidentaux modernes (Foucault 2004; Desrosières 2008b). Au-delà des modèles, c'est donc une conception du risque comme fait social, mutualisé au niveau sociétal qui préside à la mise en place des Etats Providence après 1945 (Laurent 2018; Laurent 2014). On parle alors de « sociétés assurantielles » (Foucault 2001, 387; Ewald 1986, 373; Rosanvallon 1995, 10).

- Ewald montre ainsi que se développe à partir de la fin du XIXe siècle une nouvelle forme de droit, le droit social, qui contredit la conception libérale de droits naturels ; dans cette dernière acception, les obligations sociales se limitaient à la non-interférence avec la sphère des libertés d'autrui. Dans le courant du XXe siècle, la solidarité prend forme au niveau national aux plans juridique et institutionnel avec la Sécurité Sociale, obligatoire et universelle et impose un nouveau contrat social.

- Le dernier pilier de ce modèle solidaire est le postulat de l'impossibilité d'individualiser le risque : « l'impossibilité n'est pas accidentelle. Elle a son fondement dans l'ontologie solidariste, dans la doctrine de l'être social. Sa manière de poser la précédence de la

société sur tout échange individuel, de disséminer les causes, de faire de tout une cause, interdit de jamais savoir ce que chacun doit aux autres » (Ewald 1986). L'ignorance n'est donc pas vue comme un obstacle mais comme une condition de possibilité de la solidarité. Pour citer encore Ewald : « il n'y a de risque que du point de vue d'une totalité (…) du point de vue social nous sommes des aléas : homo aleator. Ce qui fait que l'un n'est pas l'autre, c'est le fruit du hasard, des circonstances, d'une distribution hasardeuse, sédimentée, consolidée par l'histoire » (Ewald 1986), 371).

Cette ignorance constitutive et fondamentale, qui fait que le risque individuel n'est pas quantifiable alors qu'il le devient au niveau collectif reflète une communauté de destin face à l'aléa comme mal social. La régulation des risques se fait alors par la mutualisation : mutualisation des ressources qui permettront de couvrir les dommages, eux aussi mis en commun. La solidarité trouve son expression dans l'uniformité des primes, à l'exemple des premières cotisations des mutuelles de santé après-guerre (Siney-Lange 2015), 23–24). On retrouve aujourd'hui encore ce principe de solidarité dans la tarification unique de l'assurance des catastrophes naturelles, dont le montant est fixé par l'Etat.

Les récents développements technologiques apportés par le *big data* ébranlent ces piliers. Mais il ne s'agit que d'une secousse supplémentaire dans la longue histoire de la crise de l'Etat providence : au début des années 80, Rosanvallon parlait déjà des difficultés du système à financer ses ambitions et de la remise en cause idéologique du rôle de l'Etat dans la gestion des problèmes sociaux (Rosanvallon 1992). En 1995, il souligne une crise d'ordre philosophique avec « la désagrégation des principes organisateurs de la solidarité » (Rosanvallon 1995) : il voit en effet dans les débuts de la segmentation des risques une « segmentation du social » qui sape les bases du modèle de solidarité nationale de l'Etat providence. L'assurance, qui mutualise des risques homogènes, se distingue alors de la solidarité, qui impose une redistribution entre des groupes aux profils de risque différents.

Le questionnement des mécanismes de mutualisation et de solidarité induit par le *big data* s'inscrit ainsi dans une longue histoire et en radicalise certains éléments. L'approche personnalisée, renforcée par les objets connectés qui collectent des données individuelles en continu, fait ainsi apparaître l'hétérogénéité des risques au niveau le plus granulaire: le « mal social » devient attribuable à certains acteurs dont on attend qu'ils modifient leur comportement : on en vient à créer des segments à individu unique, pour reprendre une formule lancée récemment par Google (Weed 2017).

Ces attentes sont d'ailleurs fortement médiatisées par les fournisseurs d'objets connectés ; on vante ainsi la capacité des boîtiers automobiles à accroître la sécurité routière et celle des bracelets connectés à améliorer la santé. L'accent mis sur la prévention implique pourtant une redistribution de la responsabilité, qui n'est plus sociale mais individuelle et pose de façon renouvelée la question de l'équité de la tarification des produits d'assurance. La « personnalisation » de la prime sur la base de données individuelles se veut en effet aussi un gage d'équité. Les travaux effectués jusqu'ici par la chaire ont déjà remis en question le principe de l'espérance comme indicateur d'équité d'une tarification *du point de vue de l'assuré* qui, pour les accidents dont il est victime, est toujours en situation d'alea. Mais avec la multiplication des données disponibles et leur traitement parfois très opaque, quel sens donner à l'équité dans le contexte *big data*?

De ce point de vue, la législation européenne de 2012 sur l'arrêt d'usage du sexe dans la tarification est révélatrice à plus d'un titre : elle insiste tout d'abord sur la légitimité d'une tarification sur la base de la « conduite individuelle », dont la mesurabilité, on l'a vu, semble démultipliée par les objets connectés. Or la légitimité d'un facteur de risque, par le passé, relevait

d'une moyenne statistique sur un groupe d'individus appartenant à la même catégorie, mesure qui se trouve ici disqualifiée : on devrait semble-t-il tarifer au contraire sur la seule base de données individuelles, collectées en continu. Cette remarque conduit à une question de fond : assisterait-on à une transformation des modes d'appréhension du risque accompagnée et soutenue par le développement des techniques *big data* ?

Par ailleurs, le texte de la législation distingue entre le sexe, devenu facteur de tarification illégal et par là illégitime et d'autres facteurs, non cités, qui continuent eux d'être utilisables ; on pense par exemple à l'âge de l'assuré, mais qu'est-ce qui rendrait l'âge plus légitime que le sexe ? Cette contradiction indique ainsi une deuxième interrogation sous-jacente et non abordée par le texte : quels principes d'équité doivent être mis en œuvre dans la tarification des produits d'assurance, et de quelle manière, alors que l'essor du *big data* rend pléthoriques les données individuelles disponibles ? Comment répondre à cette demande de personnalisation sans créer de discrimination ?

L'assurance des catastrophes naturelles représente à cet égard un cas d'école : s'il existe en effet au sein de la CCR des modèles pour l'appréciation des risques climatiques sur la base de données géologiques et météorologiques, la tarification, elle, reste in fine très sommaire ; elle traduit une volonté de solidarité exprimée par le législateur au début des années 80, alors que les technologies ne permettaient pas beaucoup mieux que cette tarification en moyenne. Mais elle rend compte aujourd'hui d'un paradoxe : la solidarité semble résulter d'un « voile d'ignorance » - imposé historiquement par les limites du savoir de l'époque, qu'il devient de plus en plus difficile de maintenir lorsque les données existent.

Ce constat appelle une question de fond : comment gérer le risque par mutualisation et solidarité, lorsque les données deviennent pléthoriques? Données de santé ou données climatiques, la régulation impose une égalité tarifaire là où les données disponibles rendent possible une segmentation très poussée. Dans le cas des catastrophes naturelles, ce régime est cependant souvent remis en cause car, n'incitant que marginalement à la prévention via la modulation des franchises, il ne permet pas de minimiser le coût pour la collectivité dans son ensemble.[4] Doit-on pour autant basculer dans une tarification segmentée et si oui, à quel niveau de granularité ?

Au-delà du problème épistémologique sous-jacent, on voit bien que cette personnalisation va dans le sens d'une responsabilisation accrue de la personne, réduisant l'aléa de survenance à un risque comportemental, gérable par la prévention et la modification des conduites. Cette problématique se retrouve au cœur des critiques récentes de la tarification de l'assurance des catastrophes naturelles : faisant jouer à plein le principe de solidarité, ce risque est historiquement tarifé sous forme d'un pourcentage unique appliqué à la prime pour dommages aux biens, sans prise en compte de paramètres susceptibles de différencier le risque. Doit-on y voir un modèle de solidarité ou au contraire un mécanisme obsolète qui décourage la prévention et, par là, accroît les coûts pour la population dans son ensemble ? Sans préjuger de la réponse, les technologies de data science semblent en tout cas susceptibles de modifier la connaissance de nos risques et de renégocier pour nous le partage entre la part attribuée à l'aléa de survenance et celle résultant du comportement et/ou des caractéristiques intrinsèques de l'assuré.

*

Avec le paradigme d'un savoir prédictif des aléas de la personne plutôt qu'explicatif des phénomènes macros de la population, le *big data* radicalise ainsi le problème de la levée du voile d'ignorance dont Rawls avait fait le cœur d'un modèle de justice comme équité (Rawls 2005) : *l'homo aleator* semble disparaître au profit de l'homme probable, dont il conviendra de questionner la pertinence. Là encore en effet règne l'amalgame entre aléa et hétérogénéité (François and

[4] Voir notamment le projet de loi de 2012, 38 (https://www.senat.fr/leg/pjl11-491.pdf).

Frezal 2018) : il y a confusion semble-t-il entre la baisse présumée de fréquence des accidents, et la possibilité de prévenir voire de prédire l'accident individuel. La distinction entre les deux concepts permettra de préciser le rôle potentiel mais aussi les limites de la prévention dans l'assurance -prise ici au sens de régulation des risques et de l'incertain au niveau sociétal- dans un environnement *big data*.

Bibliographie

Abbott, Andrew. 1988. "Transcending General Linear Reality." *Sociological Theory* 6 (2): 169–86. https://doi.org/10.2307/202114.

Anderson, M.J. 1988. *The American census. A social history*, New Haven, Yale university press.

Baker, Wayne E. 1984. « The social structure of a national securities market », *American journal of sociology*, vol. 89, n° 4 : 775-811.

Barman, Emily. 2016. *Caring Capitalism by Emily Barman*, New York, Cambridge university press.

Beckert, Jens. 2011. « Where do prices come from? Sociological approaches to price formation », *Socio-economic review*, vol. 9, n° 4 : 757-786.

Berman, Elizabeth Popp et Daniel Hirschman. 2018. « The Sociology of Quantification: Where Are We Now? », *Contemporary Sociology*, vol. 47, n° 3 : 257-266.

Bizieux, Alban et Pierre François. 2017. « L'invention de la fonction risque : pouvoir, contre-pouvoir ? », *Working paper Chaire Pari*, vol. 10: 1-35.

Bornstein, Aaron M. 2016. "Is Artificial Intelligence Permanently Inscrutable?" *Nautilus*. September 1, 2016. http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable.

Boyer-Kassem, Thomas. 2015. "Les interprétations de la mécanique quantique | Implications philosophiques." http://www.implications-philosophiques.org/actualite/une/les-interpretations-de-la-mecanique-quantique/.

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.

Brynjolfsson, Erik, and Andrew McAfee. 2016. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. 1 edition. New York London: W. W. Norton & Company.

Carruthers, Bruce G et Wendy Nelson Espeland. 1991. « Accounting for rationality : double-entry bookkeeping and the rhetoric of rationality », *American journal of sociology*, vol. 97, n° 1 : 31-69.

Carruthers, Bruce G. 2013. « From uncertainty toward risk: the case of credit ratings », *Socio-Economic Review*, vol. 11, n° 3 : 525-551.

Chan, Cherish. 2009. « Invigorating the content in social embeddedness. An ethnography of life insurance transactions in China », *American journal of sociology*, vol. 115, n° 3 : 712-754.

Chandler, Alfred D. 1988. *La main visible des managers : une analyse historique*, Paris, Economica.

Charpentier, Arthur, Michel Denuit et Romuald Elie. 2015. « Segmentation et mutualisation, les deux faces d'une même pièce ? », *Risques*, vol. 103: 57-64.

Chen, Y et ASY Cheung. 2017. « The transparent self under big data profiling: privacy and Chinese legislation on the social credit system. », *Faculty of law research paper*, vol. 2017, n° 11.

Cusumano, Michael A., Yiorgos Mylonadis et Richard S. Rosenbloom. 1992. « Strategic Maneuvering and Mass-Market Dynamics: The Triumph of VHS over Beta », *Business History Review*, vol. 66, n° 1 : 51-94.

Daston, L.J. 1986. « The domestication of risk. Mathematical probability and insurance, 1650-1830 », *in* Lorenz Krüger, L.J Daston et M Heidelberger (éds.) *The probabilistic revolution, Vol. 1: Ideas in history*, Cambridge, MIT Press: 237-261.

David, Paul. 1985. « Clio and the economics of QWERTY », *American economic review*, vol. 75, n° 2 : 332-337.
Desrosières, Alain. 1993. La *Politique Des Grands Nombres. Histoire de La Raison Statistique*. Paris: La découverte.
Desrosières, Alain. 1998. *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, Mass: Harvard University Press.
Desrosières, Alain. 2008a. « Quételet et la sociologie quantitative. Du piédestal à l'oubli », in *Pour une sociologie historique de la quantification. L'argument statistique 1*, Paris, Presses de l'école des Mines: 239-257.
Desrosières, Alain. 2008b. *L'argument statistique. I, Pour une sociologie historique de la quantification*. Paris: Presses des mines.
Desrosières, Alain. 2014. *Prouver et Gouverner*. Paris: La decouverte. http://www.editionsladecouverte.fr/catalogue/index-Prouver_et_gouverner-9782707182494.html.
Dobbin, Frank. 2009. *Inventing equal opportunity*, Princeton, Princeton university press.
Edelman, Lauren B et Marc Suchman. 1997. « The legal environment of organizations », *Annual review of sociology*, vol. 23: 479-515.
Epstein, Joshua M. 2012. G*enerative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press.
Ewald, François. 1986. *L'Etat providence*. Paris: Grasset.
Eyal, Nir. 2014. *Hooked: How to Build Habit-Forming Products*. Penguin.
Filleule, Renaud. 2008. « La sociologie économique des prix contemporaine : quel apport théorique ? », *Année sociologique*, vol. 58, n° 2 : 383-407.
Finez, Jean. 2014. « La construction des prix à la SNCF, une socio-histoire de la tarification. De la péréquation au yield management (1938-2012) », *Revue française de sociologie*, vol. 55, n° 1 : 5-39.
Foucault, Michel. 2001. *Dits et Ecrits, tome 2 : 1976 - 1988*. Paris : Gallimard.
Foucault, Michel. 2004. *Sécurité, Territoire, Population*. Paris: Gallimard/Seuil
Fourcade, Marion et Kieran Healy. 2017. « Seeing like a market », *Socio-Economic Review*, vol. 15, n° 1 : 9-29.
François, Pierre et Sylvestre Frezal. 2018. « Instituer l'incohérence. L'amalgame aléa et hétérogénéité dans le secteur assuranciel », *Sociologie du travail*, vol. 60, n° 1.
François, Pierre. 2011. « Puissance et genèse des institutions : un cadre analytique », *in* Pierre François (éd.) *Vie et mort des institutions marchandes*, Paris, Presses de sciences po: 39-79.
Frezal, Sylvestre. 2015. « Alea et Hétérogénéité – l'Amalgame Tyrannique », Working Paper, Chaire PARI. http://www.chaire-pari.fr/wp-content/uploads/2015/11/concepts_amalgame-tyrannique-25112015.pdf
Frezal, Sylvestre. 2018. *Quand les statistiques minent la finance et la société. Risque, responsabilité et décision*. Paris : L'Harmattan.
Gervais, Pierre. 2012. « Crédit et filières marchandes au XVIIIe siècle », *Annales. Histoire, sciences sociales*, vol. 67, n° 4 : 1011-1048.
Hacking, Ian. 1975. *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge: Cambridge University Press.
Hacking, Ian. 1990. *The Taming of Chance*. Cambridge, Mass: Cambridge University Press.
Harcourt, Bernard E. 2007. *Against Prediction*. University of Chicago press.
Hay, François-Xavier. 2015. « La mutualisation est-elle soluble dans le big data ? », *Risques*, vol. 103: 25-30.
Holmes, Jessica, Jonathan Isham, Ryan Petersen et Paul M. Sommers. s. d. « Does Relationship Lending Still Matter in the Consumer Banking Sector? Evidence from the Automobile Loan Market* », *Social Science Quarterly*, vol. 88, n° 2 : 585-597.
Jarzabkowski, Paula, Rebecca Bednarek et Paul Spee. 2015. *Making a market for acts of god. The practice of risk trading in the global reinusurance industry*, Oxford, Oxford university press.

Jovanovic, Franck. 2012. « Finance in modern economic thought », *in* Karin Knorr-Cetina et Alex Preda (éds.) *The Oxford handbook of the sociology of finance*, Oxford, Oxford university press.
Karasti, Helena, Florence Millerand, Christine M. Hine et Geoffrey C. Bowker. 2016. « Knowledge infrastructures: Part I », *Science & Technology Studies*, vol. 29, n° 1.
Keynes, J. M. 1937. « The General Theory of Employment », *The Quarterly Journal of Economics*, vol. 51, n° 2 : 209-223.
Knight, Frank H. 1985. *Risk, uncertainty and profit*, Chicago, University of Chicago press.
Labbé, M. 2008. « L'arithmétique politique en Allemagne. Réceptions et polémiques », *Journal électronique d'histoire des probabilités et de la statistique*, vol. 4, n° 1.
Kuang, Cliff. 2017. "Can A.I. Be Taught to Explain Itself?" *The New York Times*, November 21, 2017, sec. Magazine. https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.
Lasry, Jean-Michel. 2015. « La rencontre choc de l'assurance et du big data », *Risques*, vol. 103: 19-24.
Laurent, Eloi. 2014. *Le bel avenir de l'Etat-providence*. Paris: Liens qui libèrent.
Laurent, Éloi. 2018. "La protection sociale : de l'incertitude au risque, de l'État Providence à l'État social-écologique." *Revue Française de Socio-Économie*, no. 20 (May): 191–94. https://doi.org/10.3917/rfse.020.0191.
Lazarus, Jeanne. 2012. « The Ambition of Credit Scoring: Forecasting Credit Failure », *Raisons politiques*, No 48, n° 4 : 103-118.
Li, Wei, Alex Oberle et Gary Dymski. 2009. « Global banking and financial services to immigrants in Canada and the US », *Journal of International Migration and Integration*, vol. 10, n° 1 : 1-29.
LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.
Martin, Olivier. 1997. « Aux origines des idées factorielles [Des théories aux méthodes statistiques] », *Histoire & Mesure*, vol. 12, n° 3 : 197-249.
Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.
Merry, Sally Engle. 2016. *The Seductions of Quantification: Measuring Human Rights, Gender Violence, and Sex Trafficking, Merry*, Chicago, Chicago university press.
Meyer, John W et Brian Rowan. 1977. « Institutionalized organizations: formal structure as myth and ceremony », *American journal of sociology*, vol. 83, n° 2 : 340-363.
Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–82. https://doi.org/10.1126/science.1199644.
Moss, David A et Daniel Carpenter. 2013. *Preventing regulatory capture*, Cambridge, Cambridge university press.
Ollivier, Carine. 2017. « L'actuaire à la croisée des chemins », *Working paper Chaire Pari*, n° 10 .
Pager, Devah et Hana Shepherd. 2008. « The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets », *Annual Review of Sociolgy*, vol. 34: 181-209.
Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, Massachusetts; London, England: Harvard University Press.
Patel, Deepak N., Estelle V. Lambert, Roseanne da Silva, Mike Greyling, Craig Nossel, Adam Noach, Wayne Derman, and Thomas Gaziano. 2010. "The Association between Medical Costs and Participation in the Vitality Health Promotion Program among 948,974 Members of a South African Health Insurance Company." *American Journal of Health Promotion: AJHP* 24 (3): 199–204. https://doi.org/10.4278/090217-QUAN-68R2.1.
Patriarca, S. 1996. *Numbers and nationhood. Writing statistics in nineteenth-century Italy*, Cambridge, Cambridge university press.

Pentland, Alex. 2014. *Social Physics: How Good Ideas Spread-the Lessons from a New Science*. New York: The Penguin Press.
Pilmis, Olivier. 2016. « Les formations d'actuaire. Une analyse sociologique », *Working paper Chaire Pari*, n° 9 : 1-25.
Rawls, John. 2005. *A Theory of Justice*. Cambridge, Massachusetts ; London, England: Harvard University Press.
Rona-Tas, Akos et Alya Guseva. 2013. « Information and consumer credit in Central and Eastern Europe », *Journal of Comparative Economics*, vol. 41, n° 2 : 420-435.
Rona-Tas, Akos et Alya Guseva. 2014. *Plastic Money: Constructing Markets for Credit Cards in Eight Postcommunist Countries*, StanfordStanford university press.
Rona-Tas, Akos et Alya Guseva. 2018. « Consumer Credit in Comparative Perspective », *Annual Review of Sociology*, vol. 44.
Rona-Tas, Akos. 2010. « The role of ratings in the subprime mortgage crisis. The art of corporate and the science of of consumer credit rating », *in* Michael Lounsbury et Paul M Hirsch (éds.) *Markets on trial. The economic sociology of the US financial crisis*, Londres, Emerald group publishing.
Rosanvallon, Pierre. 1992. *La Crise de l'Etat-providence*. Nouv. éd. Paris: Seuil.
Rosanvallon, Pierre. 1995. *La Nouvelle question sociale. Repenser l'Etat-providence*. Paris: Points.
Schuilenburg, Marc, and Rik Peeters. 2017. "Gift Politics: Exposure and Surveillance in the Anthropocene." *Crime, Law and Social Change* 68 (5): 563–78. https://doi.org/10.1007/s10611-017-9703-5.
Siney-Lange, Charlotte. 2015. *A l'Initiative Sociale* Paris: Presses du Châtelet.
Steffen, Will, Paul J. Crutzen, and John R. McNeill. 2007. "The Anthropocene: Are Humans Now Overwhelming the Great Forces of Nature?" *Ambio* 36 (8): 614–21.
Stenning, Alison, Adrian Smith, Alena Rochovská et Dariusz Świątek. s. d. « Credit, Debt, and Everyday Financial Practices: Low-Income Households in Two Postsocialist Cities », *Economic Geography*, vol. 86, n° 2 : 119-145.
Thourot, Patrick, Jean-Michel Nessi et Kessy Ametépé Folly. 2015. « Big Data et tarification de l'assurance », *Risques*, vol. 103: 46-53.
Von Plato, J. 1994. *Creating modern probability. Its mathematics, physics, and philosophy in historical perspective*, Cambridge, Cambridge university press.
Weber, Max. 1995. *Economie et société. Volume 1*, Paris, Pocket.
Weed, Keith. 2017. "The Future of Marketing? Consumer Segments of One." Think with Google. 2017. https://www.thinkwithgoogle.com/consumer-insights/unilever-consumer-marketing-segmentation/.
Yates, JoAnne. 2008. *Structuring the information age. Life insurance and technology in the twentieth Century*, Baltimore, Johns Hopkins University Press.
Zajdenweber, Daniel. 2015. « Quand la sélection augmente le risque », *Risques*, vol. 103: 54-56.
Zelizer, Viviana. 1979. *Morals and markets : the development of life insurance in the United States*, New York, Columbia university press.

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 Actuaris, 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