

CHAPITRE 2

Les traces numériques et le pouvoir d'agir des répliques

Dominique Boullier

Les sciences sociales peuvent-elles apprendre de l'apprentissage qui domine désormais la science des données (*data science*) par l'apprentissage automatique (*machine learning*) ? La mutation culturelle ne peut être ignorée, mais cela reste difficile de la saisir, de ne pas tout rejeter d'un bloc (dont les fausses promesses prédictives) ou de ne pas succomber à la tentation (« Google est trop fort, l'académie a perdu ! »). En fait, il nous paraît possible à la fois de se laisser affecter par la vague des données massives (*big data*) et de l'apprentissage automatique et de profiter de cette occasion pour mettre en évidence tout ce que les sciences sociales peuvent apporter à ces méthodes, à condition d'éclaircir ce qu'elles font et de simplifier le message. Nous proposons de montrer comment des « modèles conceptuels », au sens de Norman (1988), irriguent toutes les sciences sociales et comment une vision claire de ces divers « points de vue » équipés permet de spécifier les contributions effectives des sciences sociales dans l'exploitation des traces fournies désormais en masse par les réseaux numériques. Paradoxalement, l'informatique nous oblige à expliquer ce que les sciences sociales font dans leur « distribution d'agentivité », à la modéliser et ainsi à la rendre transférable, communicable, voire calculable.

Nous prendrons comme point de départ les propriétés reconnues des données massives pour mesurer les glissements méthodologiques auxquels nous faisons face. La qualité des données n'est plus du tout évaluée selon les critères auxquels la statistique nous avait habitués (l'exhaustivité et la

représentativité ont été remplacées par le volume et la variété, auxquels a été ajoutée la vélocité). Nous proposons de convertir cette vélocité en qualité statistique pérenne, qui serait la traçabilité, de façon à exploiter des gisements de données dans le temps et avec un grain suffisamment fin. Cette évolution d'apparence technique sera mise en perspective dans une histoire des méthodes de quantification, qui s'inspire des travaux d'Alain Desrosières, mais qui met en place trois générations différentes, toutes appuyées sur des états différents des techniques : les recensements, les sondages et désormais les traces massives. Pour éviter de penser à ces générations comme à des tables rases successives, alors qu'elles gardent chacune leur pertinence, nous montrons comment chacune de ces méthodes de quantification encapsule un certain point de vue sur le social. Loin de prétendre écraser les autres points de vue, celui des répliques rendues observables par les traces massives sur les réseaux numériques permet d'accéder à une dimension du social, celle des propagations d'entités qui traversent les esprits humains et les supports numériques. La cohérence des choix d'objets de recherche, de méthodes et d'entités agissantes constitue une condition de clarification des débats entre paradigmes au sein des sciences sociales, qui reposent en fait sur des points de vue et des méthodes différentes.

Les données massives et les 3V comme ersatz

Les critères de qualité des données massives sont souvent résumés par les « 3V » : volume, variété, vélocité. Ces 3V fournissent des équivalents, sous forme d'ersatz, des exigences classiques des sciences sociales. Notre approche n'est pas contradictoire avec l'ontologie de Kitchin et McArdle (2016), mais elle prétend être plus pragmatique et historique à la fois, plus proche des travaux de Mackenzie (2017) sur le travail des *machine learners*. La définition de la « nature » des données massives que tentent de donner Kitchin et McArdle est ici reformulée en tableau des pratiques des différentes expertises en quantification selon les époques, des institutions qui les financent et des modèles du social qu'elles incorporent pour en comprendre les limites et les potentialités. Ainsi, comme nous allons le voir, il n'est guère possible d'ajouter l'exhaustivité comme critère des données massives puisque les traces que l'on collecte en ligne n'ont pas de limites dans leur diversité (variété) et sont produites de façon dynamique (vélo-

cité), ce qui n'empêche pas le calcul grâce aux bases de données NoSQL et aux techniques d'apprentissage automatique.

Le volume et l'exhaustivité

Le volume correspond plutôt à l'exigence d'exhaustivité qui a été établie lors de ce que nous avons appelé la première génération de sciences sociales (Boullier, 2015 ; 2016 ; 2017), qui reposait sur les recensements, équipée notamment par les machines de Hollerith qui deviendront IBM. En visant l'exhaustivité, et en l'atteignant au prix d'un investissement de forme considérable (Thévenot, 1986), les États et leurs instituts de statistiques permettaient aussi aux sciences sociales d'accéder à cette entité improbable, une « société » (nationale, rappelons-le). Durkheim en fit la théorie et lui attribua une agentivité quasi toute-puissante, sans pour autant s'appuyer sur ces recensements, mais plutôt sur des registres divers qui avaient cependant la prétention eux aussi d'être exhaustifs. Le pouvoir que nous continuons à attribuer à la « société », cette agentivité spécifique de la structure sociale, doit ainsi beaucoup au dispositif qui permet de prouver qu'elle existe bien puisqu'on la dénombre, quand bien même il s'agit avant tout de la gouverner (Desrosières, 2014).

Pour les données massives, l'exhaustivité n'est que rarement atteinte, parce que les flux de nouvelles données sont récoltés en permanence et parce que, lorsqu'il s'agit du Web, cette notion n'a même plus de sens puisque ses limites sont inatteignables. Mais le volume est une condition de possibilité requise pour que l'apprentissage puisse se faire (Domingos, 2015)¹. L'exhaustivité est ainsi dégradée en volume, mais cela permet d'agréger en revanche des traces de tous types en masse, en dehors des conventions lourdes que suppose l'exhaustivité. L'acceptation de l'approximation est à la base même des méthodes de l'apprentissage automatique, qui peut valider statistiquement cette marge d'approximation et apprendre

1. L'apprentissage sera d'autant plus performant qu'il pourra réviser des corrélations trop simplistes obtenues avec des observations trop peu abondantes (phénomène de surapprentissage ou *overfitting*). Les jeux de données d'apprentissage sur lesquels sont entraînés les algorithmes, souvent avec supervision pour leur dicter ce qui est correct et ce qui ne l'est pas, doivent être différents des jeux de données d'exécution sur lesquels les algorithmes entraînés devront découvrir eux-mêmes des corrélations robustes et prédire des résultats pertinents pour le problème concerné. Cette contrainte, entre autres, explique le besoin de volume suffisant.

au fur et à mesure des nouveaux exemples ou des nouvelles sources de données. Pour les sciences sociales et pour faire convention (Eymard-Duvernay *et al.*, 2004), il conviendra de fixer un équivalent de ce volume qui se rapproche des exigences traditionnelles de l'exhaustivité, sans pour autant pouvoir les respecter notamment lorsqu'on traite du Web. Nous devons clairement faire notre deuil de l'exhaustivité sans que cela nous dispense de fixer les cadres conventionnels de toute démarche en sciences sociales traitant de traces numériques. Lorsque l'on traite de faibles volumes avec de faibles capacités de calcul, il est quasiment inévitable de retomber sur des méthodes statistiques classiques comme les régressions linéaires, qui ne pourront pas traiter des problèmes d'une échelle de complexité trop élevée. La convention à développer devra donc obliger les travaux en sciences sociales/sciences des données à énoncer explicitement l'échelle de complexité du problème d'origine, les méthodes pour la réduire (à travers des réductions de dimensions ou variables, par exemple), la quantité d'observations nécessaires pour que le calcul soit significatif (tout en tenant des capacités de calcul disponibles), les autres méthodes abandonnées en cours de route et la justification de ces abandons. Toutes ces métadonnées sur un travail de science des données se trouvent parfois explicitées dans les challenges Kaggle (compétitions en science des données) que nous avons étudiés (Boullier et El Mhamdi, 2020) et qui montrent bien le caractère empirique des pratiques des *machine learners*.

La variété et la représentativité

Le deuxième critère, la variété, est lui aussi une forme de transcription de l'exigence de représentativité qui a permis à toutes les sciences sociales de procéder par enquêtes (sondages) à base d'échantillonnage. Cette innovation statistique date des années 1930 et sa mise en lumière en 1936 par Gallup (Blondiaux, 1998) a constitué un tournant dans la quantification des sociétés en faisant émerger une deuxième génération de sciences sociales. Toutefois, de nombreuses conditions ont dû être réunies pour que ces conventions se stabilisent et encadrent les pratiques des instituts de sondage qui les réalisent ainsi que celles des médias et des marques qui les financent. Mais ce montage spécifique ne peut être transféré tel quel dans le monde numérique, notamment lorsqu'il s'agit de données native-

ment numériques (Rogers, 2013). Il est donc irréaliste de prétendre mesurer « l'opinion » sur le Web sans être capable de respecter les critères de représentativité qui ont été établis depuis plusieurs décennies (avec les intervalles de confiance notamment) et les méthodes de recueil auprès d'individus d'expressions calibrées par des questionnaires. Dès lors, le critère de la variété doit être considéré comme une version lâche de la représentativité, qui suppose que l'on accepte un niveau « suffisant » de variété pour valider un échantillon.

Pour les sciences sociales de troisième génération, qui acceptent de perdre la contrainte de représentativité telle qu'elle a été construite dans le cas des sondages, il reste à définir ce que serait cette variété. La constitution d'un ensemble de sources suffisamment variées² devrait alors répondre à quelques critères propres aux méthodes numériques et au domaine étudié. Les travaux sur l'analyse de sentiment que j'ai publiés avec Lohard en 2012 nous ont conduits à nous a conduits à considérer qu'aucune description du social-société, du social-opinion ou du social-traces ne peut être produite en généralité sur les réseaux numériques. En effet, les particularités des plateformes et leur pouvoir de formatage des traces sont trop présents, comme le montre Rogers (2013), qui propose cependant d'effectuer une réorientation (« *repurposing* ») des traces à visée commerciale pour les recentrer sur les finalités de la recherche. La prolifération des traces rend paradoxalement impossible toute prétention à une référence à un « tout » posé *a priori* ou constitué *a posteriori*. Les sciences sociales doivent accepter de ne traiter que des enjeux (« *issues* ») (Marres, 2007 ; 2017 ; Marres et Weltevrede, 2013), ou des points de focalisation d'attention, dont le numérique peut garder les traces, des traces qui seront spécifiques à chaque enjeu. Cela réduit considérablement la portée totalisante des prétentions des données massives, mais cela rend possible une certaine forme de représentativité et d'exhaustivité.

La vélocité et la traçabilité

Enfin, le dernier critère, la vélocité, ne trouve guère d'équivalent dans les sciences sociales de première et deuxième génération. On parle de vélocité à deux points de vue : d'abord la vitesse de publication des données ou

2. On parle alors de méthodes de *sourcing* (« sources d'approvisionnement ») des sites Web, par exemple pour une étude de topologie comme dans l'étude que nous avons conduite sur le Web du livre en France à partir de plus de 5 000 sites (Le Bécheuc *et al.*, 2015).

traces, car elles arrivent en temps réel puisque chaque clic est immédiatement enregistré comme trace sur les serveurs des agences de placement publicitaire par exemple, ce qui produit un flux continu ; puis, la vitesse de leur traitement et de leur mise à jour, ce qui permet de réviser en permanence les calculs et les agrégations réalisées, faites parfois avec des approximations lorsque des données sont manquantes (ce qu'on ne pouvait pas faire sans des bases de données NoSQL), et surtout d'alimenter de façon dynamique et permanente l'apprentissage des algorithmes d'apprentissage automatique. À vrai dire, ces processus dynamiques ne constituaient ni le point fort, ni la préoccupation des sciences sociales. Il était en effet essentiel de trouver avant tout à représenter les positions à un instant donné, pour montrer la force d'imposition de « la société » sur la diversité des comportements individuels ou pour montrer comment l'opinion publique évoluait à partir des expressions singulières obtenues dans les enquêtes. Certes, grâce à un suivi longitudinal des mêmes populations ou la reprise des mêmes questionnaires, il était – et il est toujours – possible de restituer l'équivalent d'une dynamique, sans jamais cependant pouvoir suivre à la trace les médiations qui permettraient de produire ces évolutions. La vitesse semble donc hors du champ des approches plus classiques.

Cependant, une branche des sciences du Web s'est emparée de cette question de vitesse à sa façon, en exploitant les traces des mêmes qui se propagent sur le Web. Il est très significatif que Kleinberg, celui-là même qui avait exporté les méthodes de la scientométrie vers l'étude de la topologie du Web, méthodes qui furent reprises par Google, se soit intéressé depuis plusieurs années (2002) à la mise au point d'un « *meme tracker* » avec Leskovec (2009). Cette dernière étude, la plus connue, a porté sur la propagation des citations durant la campagne Obama, ce qui leur permit de réaliser une visualisation spectaculaire de la focalisation de l'attention en courbes à montées et descentes très rapides (*streams and cascades*) autour de certains incidents de la campagne. Leur méthode agrège tous les types de traces que peuvent laisser ces citations, traitées comme des chaînes de caractère dont on peut trouver la trace dans tout le Web, et en produit une métrique ancrée dans le temps, au jour le jour, voire minute par minute désormais avec Twitter (l'unité de mesure étant devenue le « *tweet per second* »). La prise en compte des mêmes nous paraît prometteuse sous réserve que l'on suive aussi les transformations-traductions ou

« dérivations » de ces mêmes dans des milieux différents, car « toute existence va différant », comme le disait Tarde (1895). Il sera alors possible de parler de mémétique (Dawkins, 1976 ; Dennett, 2017 ; Shifman, 2014) et non plus seulement de viralité.

Les conditions de félicité d'une traçabilité

Dès lors, il devient possible de trouver un équivalent de la vélocité des données massives : la traçabilité. Elle devient le critère essentiel de qualité des entités que l'on peut étudier. Quelques conditions de félicité doivent être réunies pour y parvenir :

- Les traces en question doivent avoir une continuité suffisante pour qu'il soit encore possible de dire qu'il s'agit d'un même processus, sans avoir la contrainte de l'étude de la viralité (propagation à l'identique) ni le laxisme d'une intertextualité généralisée, où tout pourrait être signe et reprise de tout.
- Les traces en question doivent permettre des suivis d'associations hétérogènes, c'est-à-dire une puissance de connectivité suffisante. Pour cette raison, des traces dont le format est trop spécifique à une plateforme peu connue ne peuvent donner lieu à extension ni à suivi, quand bien même il reste nécessaire d'expérimenter, dans un premier temps, les modèles de propagation des répliques sur une plateforme précise, plus aisément maîtrisable.
- Le suivi des traces en question doit permettre de dater avec précision tous les événements, toutes les transformations et toutes les associations. Les lignes de temps (*timelines*) sont ici l'équivalent d'autres conventions, comme les points cardinaux pour la topographie ou les inégalités de richesse pour les sciences sociales de première génération.

Ces conditions rendent possible le basculement des sciences sociales vers le suivi d'éléments qui ne sont plus ni des individus, ni des groupes, ni la société, ni l'opinion. Le numérique a dissous toute certitude sur leurs statuts, qui avaient été déjà remis en cause théoriquement, puisque les plateformes comme Facebook ou Google exploitent des attributs des comptes (614 de façon standard chez Facebook, sans compter ceux que les marchands de données y ajoutent [Andreou *et al.*, 2018]). Ces éléments peuvent être des comportements élémentaires dont on peut suivre la programmation (les « j'aime » ou les *retweets*), des memes issus de la sous-culture Internet (Reddit, 9gag par exemple), des citations reprises dans

plusieurs médias et plateformes, ou même des citations scientifiques entre revues, des slogans ou visuels publicitaires ou politiques, des expressions supposées individuelles et spontanées, mais qui sont en fait répétées par des millions de comptes en quelques heures, etc. La propagation de ces éléments devient l'objet d'étude de la troisième génération de sciences sociales, car ce sont les propriétés de ces entités qui leur permettent de créer de petites différences, et par là, de circuler et d'affecter les individus et les groupes, la société et l'opinion.

Cependant, ces conditions ne sont pas nécessairement considérées comme décisives pour obtenir un accord (et faire « convention ») de la part des plateformes (les GAFA en résumé) et des marques qui font vivre tout cet écosystème par l'exploitation publicitaire de cette traçabilité. De même que la sociologie durkheimienne s'est associée de fait avec les institutions étatiques productrices de données pour produire « la société » en combinant enquêtes (des statisticiens) et registres (des administrations) (Desrosières, 1993), les instituts de sondage de Gallup et de Lazarsfeld se sont associés aux médias grands consommateurs de données sur les publics pour produire leur « opinion publique », en rapprochant ainsi « mesures d'audience » (le public des médias) et « opinion publique » (le public de la sphère publique au sens politique [Lippmann, 2008]). Les sciences sociales de troisième génération ne pourront guère faire autrement que de s'associer à ces plateformes et à ces marques pour produire la science des traces qui seraient alors traitées comme des « répliques », comme nous le proposons plus loin. Les traces produites étant dépendantes des plateformes, on ne peut guère espérer les modifier à la source. Il est en revanche possible d'exploiter les traces produites par les plateformes en les détournant de l'usage pour lequel elles avaient été conçues (selon le principe de réaffectation ou *repurposing* de Rogers [2013]). La règle veut ici qu'on ne prenne en compte aucune explication d'un autre niveau ou d'un autre monde, mais que l'on puisse comparer des vitesses de propagation, des rythmes ou des transformations éventuelles (par exemple par contamination à d'autres domaines). La différence devra se faire sur la capacité à faire voir des processus qui n'avaient jamais encore été mis en évidence, ou bien en raison des limites techniques prén numériques, ou encore des cibles adoptées par les générations précédentes de sciences sociales.

Par exemple, une étude des fausses informations, comme celle de Vosoughi *et al.*, (2018), permet de montrer que ce n'est pas leur contenu

qui génère leur viralité, mais leur score de nouveauté, car les esprits des humains sont affectés par ce qui choque, ce qui contredit ou qui surprend, et de ce fait, toute nouvelle fantaisiste et incroyable suscite une réaction, quand bien même il s'agit de montrer le caractère invraisemblable de cette nouvelle. Pour cela, les chercheurs ont dû produire un nouvel indicateur, celui de nouveauté. De même, quand on analyse les *tweets* publiés lors des émeutes de Ferguson en 2014 (Srijith *et al.*, 2017), il est possible de retrouver des histoires et des lignes narratives différentes portant ou sur le jeune, ou sur la police et qui, pour certaines, se propagent longtemps et pour d'autres, s'arrêtent assez vite. La comparaison quantifiable dans un contexte similaire « d'éléments de culture » (Dennett, 2017) assez voisins, mais différents permet d'entrer dans les propriétés de ces lignes narratives et fait comprendre ainsi leur succès relatif. Ces exemples se focalisent sur les propriétés des messages, mais d'autres travaux qui font partie intégrante de la théorie des répliques que nous proposons portent sur les propriétés des propagations. Ainsi Cheng *et al.* (2016) ont étudié les formes de propagation d'un mème célèbre (*How to be skinny*) à partir de 12 millions de messages sur Facebook. Ils ont cherché à savoir quelles étaient les conditions de la courbe d'amorçage de la propagation, laquelle permettait de prédire une propagation plus large et plus longue du mème. Ils concluent qu'un mème qui démarre faiblement a peu de chances de connaître des cascades à répétition, tout autant qu'un mème qui connaît un énorme succès dès le début de son lancement, car il semble épuiser toute l'audience potentielle d'un seul coup. En revanche, un succès modéré, mais observable dans des répliques à partir de sources différentes dès le début de la propagation constitue un facteur clé de prédiction d'une récurrence future des cascades. Cela s'explique par le fait que la diversité des milieux pénétrés permet aux différentes versions du mème de s'étendre, ce qui peut prendre du temps et faire apparaître de nouveaux pics d'attention.

Des générations aux modèles conceptuels

Le modèle des trois générations de sciences sociales est un point de départ, et il a été formulé avant tout dans la lignée des travaux d'Alain Desrosières sur la quantification qui constitue les sociétés elles-mêmes. Kitchin et McArdle (2016) proposent une distinction entre données administratives,

données d'enquête et données massives, une classification qui semble ontologique et anhistorique. Elle recoupe cependant entièrement notre proposition de générations de sciences sociales, fondée sur une histoire de la quantification inspirée des études des sciences et technologies (*sciences and technologies studies* ou STS) et qui pointe la différence de point de vue sur le social que chaque génération mobilise. Nous passons ainsi d'une vision technique des données massives et de leurs propriétés listées à l'infini à un tableau structuré des visions possibles du social encapsulées dans les méthodes.

Les différences entre générations ont pu apparaître dans la présentation précédente comme si elles étaient inscrites dans une généalogie alors que ce n'est pas du tout l'approche adoptée, quand bien même un positionnement dans une succession historique a été repéré et un héritage est bien repris et transformé. Toutefois, la coexistence contemporaine des générations doit être rappelée et, de la même façon, l'absence de hiérarchie entre les générations de sciences sociales. En requalifiant les générations de sciences sociales en termes de fréquence (haute, moyenne, basse), il est possible de faire un pas de côté vers une version plus cohérente avec l'expérience même des acteurs : la haute fréquence. Twitter constitue un monde différent par exemple de celui des modes, des opinions et des intentions de vote, qui se cristallisent à intervalles réguliers, de façon cyclique. Cependant, ils sont tous les deux sur une longueur d'onde bien différente encore des transformations démographiques qui chevauchent les générations. Les uns et les autres peuvent devenir numériques ou numérisés, mais cela ne les rend pas pour autant identiques et ce serait une erreur de les regrouper tous dans les sciences sociales computationnelles (« *computational social sciences* ») (Alvarez, 2016), alors que les phénomènes traités sont très différents.

Cependant, au-delà de ces différences de longueur d'onde, c'est une autre approche que nous voudrions développer ici, pour montrer en fait à quel point les différents types de sciences sociales que nous avons mis en évidence recoupent des oppositions classiques entre paradigmes dans les sciences sociales. Nous pensons qu'il est utile de les simplifier en autant de « modèles conceptuels », au sens de Norman (1988). Ce dernier en faisait la condition de l'appropriation des techniques, dans la mesure où tout dépliage complet des dispositifs (ici des méthodes des sciences sociales) avait tendance à agir comme un repoussoir pour les non-spécialistes, qui

se trouvent pris dans les jargons et les éventuelles controverses entre scientifiques. Alors qu'à l'inverse, tout repli complet du dispositif sur une supposée évidence des interfaces intuitives empêchait les usagers et usagères de maîtriser effectivement les outils (l'« effet boîte noire » que les sciences sociales peuvent aussi apprécier parfois et qui aboutit à qualifier abusivement certains énoncés de « science faite »). Le modèle conceptuel constituait pour Norman le bon niveau de présentation des informations pour que des prises (Bessy et Chateauraynaud, 1995) soient possibles. Présenter la diversité des sciences sociales en trois points de vue pourrait sembler abusivement réducteur à certaines personnes, et déjà trop perturbant à d'autres. Il nous semble que cela constitue pourtant le niveau des modèles conceptuels que préconisait Norman, qui redonnent prise aux non-spécialistes ainsi qu'aux spécialistes de la science des données (*data scientists*), qui les mobilisent déjà sans le savoir.

Le tableau généalogique des trois générations de sciences sociales (société, opinion et réplifications) peut être présenté en termes de postures qui président à chacune des approches. La tendance à vouloir épuiser tous les comptes rendus du monde par une seule de ces approches est fréquente et génère des guerres de tranchées et des incompréhensions, mais rien ne justifie *a priori* la disqualification de certaines des postures dès lors qu'on peut montrer leur fécondité empirique sur le terrain.

Le tableau des trois générations peut ainsi être prolongé en inscrivant chacune des générations dans une posture ou un point de vue (voir le tableau 2.1).

Les oppositions historiques en sociologie entre structure et préférences individuelles se sont rejouées sans cesse dans le combat entre sociologies d'inspiration marxiste (dont celle de Bourdieu et toutes les sociologies déterministes) et sociologies individualistes (dont celle de Boudon et l'individualisme méthodologique, mais aussi toutes celles qui ont emprunté à la théorie des jeux). Des approches de propagation ont fait leur apparition, mais plus rarement, car les influences de Leibniz ou

TABLEAU 2.1

Inscription des générations dans des postures (1)

Structure	Préférences individuelles	Propagation
Société	Opinion	Réplifications

de Deleuze d'un côté et celles de James et du pragmatisme de l'autre ont été vigoureusement combattues (par Durkheim lui-même), jusqu'à l'apparition de la théorie de l'acteur-réseau de Callon, Latour et Law (Akrich *et al.*, 2006). Le numérique donne une chance d'exister tout à fait inédite à ces approches par la traçabilité qu'il permet, quand bien même la théorie de l'acteur-réseau avait déjà posé les bases d'une méthode proprement sociologique évolutionniste, indépendamment du numérique. L'innovation scientifique et technique était d'ailleurs son terrain privilégié.

Héritages, voisinages et arbitrages

Dans la conclusion du livre *La ville-événement* (Boullier, 2010), traitant de mouvements de foule et de mouvements d'opinion, je proposais de penser les situations de changement social en tenant compte de trois dimensions qui sont autant d'influences ou de rayons imitatifs qui nous traversent et qui constituent le social. Tarde était l'inspirateur de ce modèle, jusque dans sa théorie des croyances et des désirs comme substances imitatives. Je propose donc de distinguer les héritages, les arbitrages et les voisinages.

La théorie de la société de type durkheimien repose ainsi sur une posture qui privilégie la puissance d'agir de la structure, son agentivité (la société comme structure s'impose aux individus) et le poids des héritages y est prépondérant. Bourdieu (1964) systématisera cela en parlant de « capital » et *Les héritiers* n'est pas pour rien le titre de l'ouvrage qui l'a rendu célèbre. Cette reproduction peut être reconstituée en exploitant la continuité de longues séries de données, qui prennent la forme d'ondes longues à basse fréquence.

De même, la question du voisinage est en connexion directe avec l'approche des réplifications. Observer des mouvements de foule, c'est mesurer la force de contagion de la proximité physique, comme on le voit dans les mouvements de « ola » dans les stades. Observer le *buzz* à haute fréquence, c'est mesurer la force de contagion de la proximité numérique et de la synchronisation qui (avec les notifications) produit sur le plan temporel l'équivalent de la proximité spatiale. Il s'agit toujours de plusieurs voisinages, et non de celle dont on parle par exemple dans des quartiers urbains. Une modélisation d'effets de voisinage comme celle de Schelling (1960) repose quant à elle sur la simplification extrême des capacités de

décision des entités modélisées : elles quittent leur quartier dès lors qu'un seuil de similarité (noir/blanc) est franchi, sans réellement penser aux propagations entre voisins. Elle relève plutôt de la deuxième génération, c'est-à-dire des préférences individuelles et des capacités de calcul des agents.

Ces deux influences, héritages et voisinages sont en conflit, elles génèrent de l'hésitation, disait Tarde (1895), et dans ce conflit se loge toute la dynamique sociale, qui permet d'observer du changement et non une pure reproduction, mais du changement qui s'installe dans la durée et finit par faire structure et non une agitation permanente, quand bien même la frénésie des *tweets* pourrait le laisser penser.

La question archi-rebattue de « changement et continuité » reviendrait donc en force ici, mais elle n'a guère d'intérêt, si ce n'est de nous permettre d'observer qu'elle provient, d'un point de vue pragmatiste, de véritables luttes d'influence, dont on ne peut comprendre le résultat qu'après coup. En effet, c'est l'hésitation qui doit être aussi formalisée. C'est pourquoi je propose de prendre en compte, au même niveau, les enjeux de l'arbitrage qui s'appuient souvent sur un modèle de l'agent calculateur, idéal de l'économie fétichisée et modélisée avec théorie des jeux à l'appui, mais qui reposent aussi sur des dynamiques de désirs. Effectivement, ce sont les désirs (et donc les peurs) et l'énergie qui est investie dans cet arbitrage qui sont en jeu, comme l'avait dit Tarde (1895). L'arbitrage ne renvoie pas à une prise de décision, mais plutôt à une compétition entre désirs qui peuvent se mesurer et qui exercent alors leur influence sur nos comportements. Le concept de « marché » cherche à saisir ces moments d'arbitrage, ces préférences individuelles, par la mécanique des prix et plus exactement de la monnaie (et non de la valeur) (Orléan, 2011). Cet arbitrage tend à devenir autoréférentiel (pour l'individu, pour la société comme pour la théorie) en particulier dans le capitalisme financier numérique qui est le nôtre, mais les forces des voisinages et des héritages jouent comme des forces de rappel constantes.

TABLEAU 2.2

Inscription des générations dans des postures (2)

Structure	Préférences individuelles	Propagation
Société	Opinion	Répliques
Héritages	Arbitrages	Voisinages

Dès lors, les postures théoriques ne sont plus incompatibles ni rivales, si l'on admet qu'elles rendent compte de processus d'influences différents, sans espérer rendre compte d'un tout ou d'une combinaison de déterminations. Elles ne sont plus seulement issues des méthodes de quantification et donc d'objectivation mises en œuvre à des époques différentes.

Si le numérique doit permettre un saut, c'est sur le plan du calcul, notamment parce que des degrés de structure, des degrés de préférences individuelles ou de réplifications peuvent être calculés et donc discutés et combinés alors qu'une approche par « essence structurelle » (et non structurale) est obligatoirement binaire (présence ou non). Dans les visions totalisantes, l'agentivité n'est pas vraiment distribuée, elle est entièrement confiée, pour les besoins de l'analyse, à la structure, aux préférences individuelles ou aux réplifications. Le potentiel de classification, d'expérimentation et de validation de ces classes bien connu des informaticiens prend alors tout son sens. Les héritages (et donc les structures) se calculent en degrés d'irréversibilité, ce qui permet de comprendre les effondrements rapides si l'on choisit les bons indicateurs. Le marché et l'arbitrage deviennent des degrés de désirabilité et ni l'utilité ni le travail ne constituent leur référence externe, mais le jeu des attentes réciproques, dont la rivalité mimétique de Girard (1979), pouvait donner une idée (une métrique comme celle de la volatilité d'un titre en bourse peut la rendre calculable). Les réplifications relèvent, elles, de degrés d'immunité (Sloterdijk, 2005) puisque c'est dans la perméabilité des frontières que le voisinage vient perturber les effets des héritages. Selon la porosité plus ou moins grande de ces frontières (notre réceptivité), la circulation des entités (idées, slogans, mêmes) qui nous voient et nous traversent sera plus ou moins facilitée.

La haute fréquence des réseaux numériques demande une théorie des réplifications

L'absence de méthodes empiriques fiables pour certaines approches permet seule d'expliquer l'ignorance ou la prétention hégémonique de chacune des approches. Le seul porteur de l'approche des traces et des propagations, Gabriel Tarde, pourtant quantitativiste qualifié, n'a jamais bénéficié d'une reconnaissance universitaire comparable à celle de Durkheim ou de Weber. Ce n'est pourtant pas le numérique qui fait

émerger des entités inédites, il se contente de rendre visibles et calculables ces entités (les répliques) qui faisaient partie de l'expérience ordinaire (comme les conversations et les rumeurs) et qui avaient été conceptualisées par Tarde (1895) dans sa théorie de l'imitation (qui comprend aussi opposition et invention et ne se limite donc pas à une simple duplication durant la propagation). Ce que nous avons appelé une troisième génération de sciences sociales existait donc potentiellement, mais comme l'ont montré les études des sciences et technologies (STS) et la théorie des conventions, tant que les dispositifs de mesure n'étaient pas disponibles ni étendus pour rendre commensurables les entités en question, elles demeuraient dans les limbes des sciences sociales. L'ère numérique et la traçabilité généralisée qu'elle permet (et qui s'étend désormais aussi aux objets les plus ordinaires) fournit une nouvelle métrique centrée sur les propagations et permet aux sciences sociales, si elles s'en emparent, de dégager un nouveau point de vue sur le social, celui des répliques. Certes, toutes les approches, tous les points de vue et tous les modèles conceptuels peuvent prétendre exploiter la numérisation générale du monde. Toutefois, il serait surprenant que celle qui bénéficie le plus de la visibilité nouvelle fournie à « ses » entités (ce qui circule, les idées et les rumeurs qui nous traversent et nous font agir) ne renforce pas une science sociale spécifique, celle des répliques, encore à inventer.

Un tableau synchronique des approches en question permet de restituer leurs différences et de leur donner toute leur place, sans prétendre à une explication totale à partir d'un point de vue unique. Chacune d'elles est un point de vue sur le monde, sachant qu'il reste impossible de saisir tous les points de vue à la fois (avec toutes les limites de l'analogie, la physique quantique procède ainsi). Ce point de vue dépend entièrement des instruments qu'il mobilise et qui cadrent le monde tout en permettant sa saisie en même temps (Latour, 1990). Néanmoins, ces points de vue se focalisent sur des entités particulières qui ont dans chaque cas un statut privilégié, un pouvoir d'agir, une agentivité, d'emblée privilégiée et distribuée inégalement entre les entités. La focalisation sur les « structures », associée à l'exhaustivité des recensements, diffère nettement de la focalisation sur les « préférences individuelles » associée à la représentativité des sondages. Les résultats sont mis à contribution avant tout dans le marketing, dans les études d'opinion publique, mais aussi dans toutes les grandes enquêtes d'attitude utilisées par les économistes, par exemple sur

TABEAU 2.3

Trois points de vue en sciences sociales

Point de vue	Structure	Préférences individuelles	Réplications
Concepts clés	Système, reproduction, patrons	Choix rationnel, prise de décision, théorie des jeux	Événement, incertitude, propagation, chaos, crise
Distribution de l'agentivité (du pouvoir d'agir)	Collectifs, agrégats, communautés, réseaux	Individus, acteurs, stratégies, influenceurs	Entités circulantes, ondes, mêmes
Sources des croyances et des désirs	Héritage	Arbitrage	Voisinage
Longueur d'onde	Ondes longues	Cycles	Ondes à haute fréquence
Analyse de réseaux	Structure du réseau	Nœuds	Flux

le sentiment de bonheur ou de confiance en l'avenir, etc. (par exemple, Algan et Cahuc, 2010). Ces deux entités (structure sociale et préférences individuelles agrégées sous le nom de marché) sont elles-mêmes bien différentes des entités circulantes (Le Béhec et Boullier, 2014), qui peuvent être des messages, des biens ou des êtres qui ont leurs propriétés spécifiques analysées parfois en sémiotique, mais traitées avant tout dans la mémétique.

Le travail des sciences sociales pour penser leur propre activité se trouve ainsi éclairé par la contrainte d'explication propre au numérique. En cherchant à savoir ce que les sciences sociales peuvent réellement faire des traces du numérique, de ces traces nativement numériques rencontrées sur le Web et sur les réseaux sociaux, nous avons bien compris que les problèmes soulevés n'étaient pas du même ordre que ceux posés par les données issues d'autres générations de méthodes de quantification, quand bien même elles sont numérisées (recensements et sondages sont aussi numériques, évidemment). D'autres entités s'imposent dès lors que l'on peut les tracer et les mesurer (quand bien même elles étaient là bien avant !). La priorité nous semble donc de contribuer à construire cette théorie des répliques, des phénomènes de haute fréquence qui résultent

de voisinages et de circulations d'entités qui demandent à être caractérisées. Néanmoins, en plaçant cette nouvelle piste en perspective historique, nous ne disqualifions aucune des générations précédentes, aucun des modèles conceptuels qui ont fondé les sciences sociales (les structures et les préférences). Et chacune d'elles s'emparera des outils de calcul de l'apprentissage automatique à sa façon, sans qu'il soit nécessaire de chercher à les unifier. Analyser les réseaux de façon structurale ou sur le mode égocentré en privilégiant les nœuds et leur agentivité ne doit pas faire oublier ce qui circule et garde son pouvoir d'agir (et que certaines analyses de réseaux ont eu tendance à sous-estimer). Les méthodes numériques seront donc très différentes selon la distribution d'agentivité que l'on fait dès le départ sans espoir de reconstituer un quelconque « tout » (Latour *et al.*, 2012).

Références

- AKRICH, Madeleine, Michel CALLON et Bruno LATOUR, *Sociologie de la traduction : textes fondateurs*, Presses des Mines, 2006.
- ALGAN, Yann et Pierre CAHUC, « Inherited Trust and Growth », *American Economic Review*, vol. 100, n° 5, 2010, p. 2060-2092.
- ALVAREZ, R. Michael (dir.), *Computational Social Sciences. Discovery and Prediction*, Cambridge University Press, 2016.
- ANDREOU, Athanasios *et al.*, « Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations », dans *Network and Distributed Systems Security (NDSS) Symposium 2018*, San Diego, 2018, p. 1-15.
- BESSY, Christian et Francis CHATEAURAYNAUD, *Experts et faussaires. Une sociologie de la perception*, Métailié, 1995.
- BLONDIAUX, Loïc, *La fabrique de l'opinion. Une histoire sociale des sondages*, Éditions du Seuil, 1998.
- BOULLIER, Dominique, « Big data challenges for social sciences: From society and opinion to replications », *International Sociological Association esymposium*, vol. 7, n° 2, 2017.
- BOULLIER, Dominique, *Sociologie du numérique*, Armand Colin, [2016] 2019.
- BOULLIER, Dominique, « Les sciences sociales face aux traces du Big Data. Société, opinion ou vibrations ? », *Revue française de Science politique*, vol. 65, n° 5-6, 2015, p. 805-828.
- BOULLIER, Dominique, *La ville-événement. Foules et publics urbains*, Presses universitaires de France, 2010.
- BOULLIER, Dominique et El Mahdi EL MHAMDI, « Des modèles aux pratiques: le Machine Learning à l'épreuve de la complexité », *Revue d'anthropologie des connaissances*. Commentaire sur le texte source: 14-1, 2020, en ligne. <http://journals.openedition.org/rac/4260>.

- BOULLIER, Dominique et Audrey LOHARD, *Opinion mining et Sentiment analysis: méthodes et outils*, Open Editions Press, 2012.
- BOURDIEU, Pierre et Jean-Claude PASSERON, *Les héritiers: les étudiants et la culture*, Éditions de Minuit, 1964.
- CHENG, Justin et al., « Do cascades recur? », dans WWW, 16 *Proceedings of the 25th International Conference on World Wide Web*, Montréal, 2016, p. 671-681.
- DAWKINS, Richard, *The Selfish Gene*, Oxford University Press, 1976.
- DENNETT, Daniel, *From Bacteria to Bach and Back. The Evolution of Minds*, Penguin Books, 2017.
- DESROSIÈRES, Alain, *Prouver et gouverner. Une analyse politique des statistiques publiques*, La Découverte, 2014.
- DESROSIÈRES, Alain, *La politique des grands nombres. Histoire de la raison statistique*, La Découverte, 1993.
- DOMINGOS, Pedro, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, 2015.
- EYMARD-DUVERNAY, François et al., « L'économie des conventions ou le temps de la réunification dans les sciences sociales », *Problèmes économiques*, n° 2838, 2004, p. 1-8.
- GIRARD, René, *La violence et le sacré*, Éditions Grasset, 1979 [1972].
- KITCHIN, Rob et Gavin MCARDLE, « What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets », *Big Data & Society*, vol. 3, n° 1, 2016, p. 1-10.
- KLEINBERG, Jon, « Bursty and hierarchical structure in streams », dans *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, 2002, p. 91-101.
- LATOUR, Bruno, *La science en action*, La Découverte, 1990 [1987].
- LATOUR, Bruno et al., « The whole is always smaller than its parts: A digital test of Gabriel Tarde's monads », *British Journal of Sociology*, vol. 63, n° 4, 2012, p. 590-615.
- LE BÉCHEC, Mariannig et Dominique BOULLIER, « Communautés imaginées et signes transposables sur un "Web territorial" », *Études de communication*, n° 42, 2014, p. 113-125.
- LE BÉCHEC, Mariannig, Dominique BOULLIER et Maxime CRÉPEL, *Le livre-échange. Vies du livre et pratiques des lecteurs*, C & F Éditions, 2018.
- LESKOVEC, Jure, Lars BACKSTROM et Jon KLEINBERG, « Meme-tracking and the dynamics of the news cycle », dans *KDD '09 Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 2009, p. 497-506.
- LIPPMANN, Walter, *Le public fantôme*, Démopolis, 2008 [1925].
- MACKENZIE, Adrian, *Machine Learners: Archaeology of a Data Practice*, The MIT Press, 2017.
- MARRES, Noortje, *Digital Sociology*, Polity Press, 2017.

- MARRES, Noortje, « The issues deserve more credit: Pragmatist contributions to the study of public involvement in controversy », *Social Studies of Science*, vol. 37, n° 5, 2007, p. 759-778.
- MARRES, Noortje et Esther WELTEVREDE, « Scraping the social? Issues in live social research », *Journal of Cultural Economy*, vol. 6, n° 3, 2013, p. 313-335.
- NORMAN, Donald A., *The Design of Everyday Things*, Basic Books, 1988.
- ORLÉAN, André, *L'empire de la valeur*, Éditions du Seuil, 2011.
- ROGERS, Richard, *Digital Methods*, The MIT Press, 2013.
- SCHELLING, Thomas C., *The Strategy of Conflict*, Harvard University Press, 1960.
- SHIFMAN, Limor, *Memes in Digital Culture*, The MIT Press, 2014.
- SLOTERDIJK, Peter, *Sphères III - Écumes*, Maren Sell Éditeurs, 2005.
- SRIJITH, P. K. et al., « Sub-story detection in Twitter with hierarchical Dirichlet processes », *Information Processing & Management*, vol. 53, n° 4, 2017, p. 989-1003.
- TARDE, Gabriel, *La logique sociale*, Félix Alcan, 1895.
- THÉVENOT, Laurent, « Les investissements de forme », dans Laurent THÉVENOT (dir.) *Conventions économiques*, Presses universitaires de France et Centre d'études de l'emploi, Paris, 1986, p. 21-71.
- VOSOUGHI, Soroush, Deb ROY et Sinan ARAL, « The spread of true and false news online », *Science*, vol. 359, n° 6380, 2018, p. 1146-1151.