Les services de renseignement à la porte du Big Data

Marc Vidal – 6 octobre 2015

A – le contexte

La loi sur le Renseignement, adoptée par le Parlement, a été validée pour l'essentiel par le Conseil Constitutionnel le 23 juillet 2015.

Les inquiétudes que soulève cette loi sont nombreuses. Mon but n'est pas ici de reprendre l'argumentation des opposants mais de me focaliser sur un point particulier, celui des « *boîtes noires* », c'est à dire du captage des données de connexion dans un but d'analyse automatique.

Il convient toutefois de remarquer que cette loi ne traite pas des activités des services de renseignements mais uniquement des interceptions de sécurité (IS) élargies aux nouvelles possibilités techniques qu'ouvrent les réseaux et la géolocalisation.

Le Renseignement humain, les pratiques offensives ne font l'objet d'aucun encadrement législatif précis.

Le flou entourant les véritables missions s'étend en réalité à tout le champ d'action des services. Il n'existe aucun accès indépendant aux fichiers créés et manipulés par les services : la CNIL (Commission Nationale Informatique et Liberté) n'a aucun droit de regard sur les fichiers administratifs des services, même si elle peut être consultée au moment de la création.

On n'en sait guère plus sur les échanges de fichiers de données entre services de renseignements étrangers (et notamment l'accord Lustre qui existerait entre la DGSE et le NSA). C'est une pratique connue des Services d'utiliser des opérateurs étrangers, donc non soumis à un droit national, pour réunir des informations non légalement accessibles.

A défaut de contrôle *à priori*, il existe une commission chargée de contrôle *à postériori*. La DPR (Délégation Parlementaire au Renseignement) crée en octobre 2007 est composée de parlementaires dont la mission est d'évaluer les pratiques des services de renseignement. (Rappelons que la constitution de 2008 charge le Parlement de « contrôler l'action du gouvernement »)

En réalité cette Commission n'a pas les moyens de réaliser de véritables investigations. Elle agit plus comme un porte-parole des Services que comme un Contrôleur[1].

Dans son rapport de 2014, La DPR a écrit une phrase prophétique : « La loi devra donc affirmer ce principe : jamais nos concitoyens ne pourront faire l'objet d'un espionnage massif ».
Mais le prophète a tout faux. La loi Renseignement qui vient d'être votée instaure au contraire une possibilité légale de branchement direct sur la totalité des données des Fournisseurs d'Accès Internet (FAI) ainsi que sur celles des opérateurs de téléphonie.

Cette loi « relative au renseignement » est la première tentative en France d'encadrer l'activité des services de renseignements mais, répétons-le, uniquement du point de vue des dispositifs techniques. Un de ses articles (l'article L851-3) autorise des « traitements automatisés » sur les « informations et documents » présents sur les « réseaux ».

Bien sûr les observateurs se sont beaucoup interrogés sur ce que pouvaient être les traitements en question.

1 Les rapports de la DPR sont disponibles en ligne.

La réponse est peut-être à chercher du côté du *Big Data*.
On regroupe sous ce nom un ensemble de techniques d'analyse qui s'appliquent spécifiquement aux forts volumes de données, tels qu'on peut les obtenir par le captage des flux numériques sur les routeurs et les câbles assurant le trafic Internet, ainsi que sur les systèmes téléphoniques (GSM, GPS, réseau 3G, réseau 4G).

Boîtes noires

Ce dispositif de captage, uniquement autorisé dans le cadre de la lutte contre le terrorisme, est connu sous le nom journalistique de « *boîte noire* ».
On peut comparer la totalité des données à une *meule de foin* qui contient quelque part une aiguille.

Dans ce captage, on ne s'intéresse théoriquement qu'aux méta-données[2]. Si le traitement fait apparaître des « données de connexion » suspectes, les services ont le droit de descendre jusqu'au contenu des communications, puis dans une dernière étape à l'internaute concerné (*l'aiguille* dans *la meule de foin*).

Je propose d'appeler ce processus « le passage de l'Internet des données à l'Internet des personnes ».

Les données captées par les boîtes noires peuvent être conservées quatre ans.

Tous les aspects techniques de cette procédure sont sous le contrôle du pouvoir exécutif (le cabinet du Premier Ministre) via des avis d'une instance administrative, la CNCTR (Commission nationale de contrôle des techniques de renseignement).

A ce niveau de la procédure on repasse dans du connu : surveillance du contenu des communications de la personne visée et de ses proches, renseignement humain, actions offensives (recrutement d'indicateurs, désinformation, manipulation etc..) ou signalement à la justice.

Bien que la loi reste floue sur bien des aspects, une chose est assez évidente : présentée comme elle l'est, la procédure ne sert à rien.

Lorsqu'on inspecte des données de qualité douteuse, il est à peu près certain qu'on va obtenir des *faux positifs*, c'est à dire des individus que le logiciel a déclaré « *suspects* » mais que l'enquête va dédouaner de toute intention terroriste.

Comme les données captées sont extrêmement nombreuses, les faux positifs seront très nombreux aussi et, en réalité, très au-delà des capacités d'investigation des Services.

Nous allons voir maintenant si on peut s'éclaircir les idées avec ce qu'on sait sur les techniques de *Big Data.* Ellesi soulèvent depuis quelques années un intérêt croissant dans le monde scientifique. Ce sont des techniques très liées à l'économie, à l'entreprise, au marketing. Mais elles ne sont pas inintéressantes dans le domaine de politiques publiques ou du renseignement.

Dans le reste de ce texte, nous allons introduire les principales notions du *Big Data* dans l'optique de ce qui nous intéresse ici. Nous ferons un rapide panorama des outils qui existent avec une parenthèse particulière sur l'analyse du texte. On s'interrogera ensuite sur les fichiers sur lesquels peuvent travailler ces outils.
Dans une dernière partie, on s'attachera à relever quelques points qui font problème dans la loi sur le renseignement.

2 Dans le texte de loi, on parle aussi d' « informations et documents ».

B - Les techniques Big Data

B1 - Principaux concepts

La numérisation de notre environnement produit un grand nombre de données numériques[3], parfois à l'initiative d'êtres humains (un courrier, un chiffre de vente...), parfois par le simple fonctionnement d'un capteur (la vitesse d'une voiture, une consommation électrique...).

Le *Big Data* recouvre tous les types de données : celles qui sont sous forme d'images (par exemple photos de vacances, images de vidéo-surveillance), celles qui sont sous forme de chiffres, celles qui sont sous forme de texte (discussion téléphonique, articles, courriers individuels...).

Pour le moment le traitement automatique d'images est à la marge du Big Data. Elles sont manipulées comme une boîte noire éventuellement avec des méta-données sans traitement de leur structure interne.

Ces données peuvent être directement reliées à un être humain, comme par exemple une discussion par SMS ou une géolocalisation.
D'autres données peuvent être extraites de l'environnement, comme par exemple le temps qu'il fait.

Il est intéressant d'unifier les deux types de données. Si on dispose par exemple d'une description exhaustive des déplacements dans une ville en fonction de l'heure et du temps qu'il fait, on pourra agir pour fluidifier le trafic.
Dans ce cas précis, le *Big Data* sert à prévoir puis gérer (*monitorer*) une situation en temps réel. Ce n'est pas sa seule utilisation mais on sent bien que les données peuvent aider la gestion publique y compris en situation d'urgence.

Méta-données

La législation se veut prudente lorsqu'on passe du niveau des grands tableaux de données à des données individuelles, anonymisées ou non.

Les textes de loi font souvent une distinction entre les méta-données (appelées données de connexion[4]) et les données proprement dites, les premières décrivant les secondes.

Ainsi dans un courrier, l'heure d'envoi, les destinataires sont des méta-données encapsulées dans le message lui-même. Mais la nature de fichiers attachés, la fréquence d'envoi, voire le nombre de mots sont aussi des méta-données qu'un analyste peut extraire sans lire les messages eux-mêmes.

Le recueil des méta-données n'est absolument pas anodin et en dit parfois plus que les données elles-mêmes. Les exemples peuvent être innombrables : un appel au Samu (problème de santé ?), une interruption brutale d'envois de messages entre deux personnes (brouillés ?), une connexion à un site de rencontre pornographique (habitude sexuelle ?), le contact avec une banque suisse (évasion fiscale ?), l'écoute d'une télévision (qui émet de quel pays ?) ou d'une émission à une heure donnée (qui parle de quel sujet?).

On peut facilement construire des graphes de relations à partir des échanges transitant par les

---

3 Pour une vision générale sur le Big Data, voir :
- *Big data : nouvelles partitions de l'information : actes du séminaire IST Inria*, octobre 2014 -- ouvrage coordonné par Lisette Calderan, Pascale Laurent, Hélène Lowinger... [et al.]
- *Big Data, penser l'homme et le monde autrement* -- Gilles Babinet

4 Théoriquement les données de connexion sont celles indispensables aux intermédiaires (FAI pour l'essentiel) pour rendre le service demandé mais le but des traitements automatisés est précisément de produire de nouvelles données.

réseaux. Il peut être aisé d'enrichir ensuite ces informations (une loge de francs-maçons ?).
On est là dans un terrain bien balisé. Cela fait une éternité que la psychosociologie construit et analyse ce type de graphe. On sait facilement *segmenter une population* à partir de données de connexions.
Les seules connexions de téléphonie mobile peuvent servir à des buts de profilage de locuteurs, de décisions relatives à l'organisation des déplacements ou aider à des choix spatiaux en situation de catastrophe.

L'anonymat : au pseudonyme prés

Dans le cadre du respect de la vie privée, les données du Big Data sont supposées être anonymes. Cela n'empêche nullement une dés-anonymisation.

Les utilisateurs de données numériques laissent des traces qui peuvent permettre un repérage dans le flot des données. On parle de « signature numérique ».
Pour donner un exemple de signature numérique, on peut géo localiser un trajet quotidien, sauf le week-end, entre l'adresse A (vraisemblablement le domicile) et une adresse B (vraisemblablement le lieu de travail).

Ces signatures numériques peuvent être plus ou moins précises et aller d'indices ténus (un style d'écriture) à des indices plus robustes (le nom qui apparaît dans l'adresse du courriel du destinataire, une liste de contact...).

Il est bien sûr très intéressant de repérer et croiser toutes ces signatures et le Big Data permet ce type de manipulation.

Inversement si on connaît quatre données personnelles, on peut accéder à toutes les informations concernant la personne en question parmi des millions d'enregistrements[5].

En réalité, il existe une donnée qui est indirectement nominative et qui est l'adresse IP de « localisation de l'équipement terminal », donnée dont la collecte est autorisée. Cette adresse permet d'accéder rapidement à l'identité de la personne abonnée (c'est le fournisseur d'accès Internet qui génère l'adresse IP pour son client, en général sous forme statique). Plutôt que de parler de données anonyme, il convient donc de parler de données pseudonyme[6]

B2 - Les outils d'analyse des données : prévoir sans comprendre

La nouveauté de la situation actuelle dans le monde numérique tient dans l'avalanche des données[7], avec un nombre toujours plus important d'échanges sur les réseaux et surtout, un nombre toujours plus important de capteurs. Ces capteurs sont susceptibles d'aider les spécialistes d'un métier. Ils peuvent être très éloignés d'un comportement individuel (par exemple l'usure d'une pièce d'avion) ou au contraire directement liés à la vie d'un individu (par exemple la courbe de consommation électrique).

Les méthodes pour traiter de vastes ensembles de données sont connues depuis les années 60. Les mathématiques, les statistiques, la Recherche Opérationnelle ont multipliés les modèles et les programmes informatiques : analyses multifactorielles, modèle de classifications ascendantes ou

5 Cette étude est parue dans la revue « scientific report ». Source : le supplément spécial du *Monde du 30 septembre 2015* consacré au Big Data.

6 Le terme est dans le projet de règlement européen sur la protection des données personnelles (en cours d'adoption). Origine de l'info : le *mémoire en réplique contre la loi relative au renseignement*.

7 On convient de parler de *Big Data* lorsqu'il y a, à la fois un fort Volume de données, une grande Variété des données et une forte Vélocité dans leur renouvellement (les trois V). Les données sont récupérées telles quelles avec leur incomplétudes, incohérences et erreurs.

descendantes, modèles d'optimisation de processus (le PERT étant sans doute le plus connu)... Certains de ces modèles se focalisaient sur l'évolution d'une série dans le temps (ARMA), ce qui peut être très intéressant pour la décision stratégique.

Il faut bien dire que ces spécialités n'intéressaient pas grand monde.

Ce n'était pas le cas de la fouille dans les bases de données (*data-mining*) qui font en petit ce que le *Big Data* peut faire en gros. Ces techniques sont plutôt mise en œuvre dans des buts d'intelligence économique.

D'autres chercheurs ont définis des outils qui prévoient une situation à partir d'un apprentissage sur des données du passé (*machine learning*)[8].

Chacun de ces outils est potentiellement un de ces « traitements automatisés », cités dans l'article 651-3 de la loi relative au renseignement.

La palette des effets peut être très large : simplifier les données, visualiser les données, faire apparaître des « signaux faibles », voire des corrélations complètement inattendues. Il peut s'agir aussi de trier les données pour ne retenir que les enregistrements qui correspondent à un modèle (*pattern*), par exemple de communication terroriste.

En réalité, tout pose question sur les effets sociaux que peuvent avoir ces outils.

L'extraction du « terroriste présumé » peut se faire par exemple par affinements successifs. Pour fixer les idées, imaginons une classification descendante. Elle peut regrouper « ceux qui ne votent pas » (info obtenue par géolocalisation) puis plus détail « ceux qui ne votent pas et qui regardent une télévision arabe » puis plus détail »ceux qui ne votent pas, qui regardent une télévision arabe et qui vont à la mosquée ».

Ce sont donc des *segments de population* qui peuvent être retournés par les algorithmes. Dans ces segment plus ou moins précis et plus ou moins suspects, il devient important d'analyser plus en détail les *tendances* qui apparaissent.

Dans le meilleur des cas, la tendance permet *d'extrapoler* de de prévoir un danger auquel on n'aurait pas pensé spontanément.

Cette analyse des tendances suppose inévitablement une conservation des données.

D'ailleurs, la loi sur le renseignement prévoie que les données de connexion recueillies pourront être conservées quatre ans. La justification est donnée dans les commentaires du gouvernement sur la loi[9] : « reconstituer les parcours sur une longue période », « repérer les déplacements à l'étranger », « reconstituer des réseaux (NB: terroristes) ».

Cette conservation très vaste des données de la population n'exclue pas la nécessité de fichiers de travail indispensables aux traitements automatisés et qui échappent à toute déclaration ou contrôle (ce point a fait l'objet de critique de la part de la CNIL[10]).

Ces jeux de données sont également indispensables pour pouvoir faire fonctionner les logiciels de *machine learning*, qui se nourrissent de données passées pour générer un programme qui sera capable de prévoir.

On est typiquement là dans une situation où on *peut prévoir sans comprendre*, ce qui est d'ailleurs un champ important du *Big Data*.

---

8 Sur l'apprentissage automatique, voir *Big data et machine learning : manuel du data scientist* -- Pirmin Lemberger,... Marc Batty,... Médéric Morel,...

9 *Observations du gouvernement sur la loi relative au renseignement* (disponible sur le site de l'Assemblée Nationale)

10 La Commission Nationale Informatique et Libertés a produit un avis, disponible en ligne : *Délibération n02015-078 du 5 mars 2015 portant avis sur un projet de loi relatif au renseignement*.

La prise de décision en urgence (le *monitoring*) suppose une surveillance en temps réel des données. L'article 851-2 en parle uniquement comme élément de surveillance d'un suspect de terrorisme, c'est à dire comme un travail policier classique.

Gardons à l'esprit que le monitoring des réseaux peut avoir un véritable intérêt de politique publique. Dans le secteur de la santé par exemple, on peut anticiper les effets de propagation d'une épidémie. Google en a fait la démonstration pour les épidémies de grippe. De façon significative, Google s'est d'ailleurs trompé dans l'ampleur de l'épidémie.

B3 - Une voie royale : le texte pour sonder les opinions

L'article 853-3 de la loi renseignement parle « de détecter des connexions susceptibles de révéler une menace terroriste ». Le traitement automatisé doit retourner « des identifiants signalés » et apparemment des « informations et documents »[11].

La formulation est étrange mais au moins, on peut être sûr qu'il y a toute légalité à récupérer les données offertes publiquement par les internautes : blogs, commentaires, réseaux sociaux, thématique des photos etc. Evidemment cela suppose de passer d'une donnée de connexion (au compte twitter par exemple) au contenu textuel public, et ceci en gardant l'anonymat de l'internaute.

On entre ainsi dans l'**analyse des données textuelles** (*text-mining*) émises sous forme écrite ou orale. Les traitements du langage naturel intéressent beaucoup les services marketing des entreprises, car cela permet de savoir ce qui se dit sur un produit ou un service. L'analyse du corpus des *hotlines* se fait aussi parfois par ce moyen.

Un corpus textuel peut donner lieu à toutes sortes de traitement : résumé automatique, question/réponse, recherche d'une thématique, confrontation (*mapping*) avec une thématique prédéfinie.

On peut repérer par exemple les phrases qui ont rapport à la fraude fiscale, à partir d'un premier travail de segmentation de la population (par exemple la population en relation avec un fiscaliste spécialisé).

Bien au-delà du texte lui-même, des outils spécialisés permettent de repérer des opinions, des sentiments, des évolutions.

On peut ainsi anticiper comment sera reçu un nouveau produit, un film, une émission.

Dans le domaine des politiques publiques, un traitement automatisé remplace avantageusement une enquête d'opinion. Il s'agit donc d'un puissant outil d'aide à la décision ou à la communication gouvernementale.

Le texte peut aussi révéler un *sentiment prédisposant à l'action*, ce qui est le but recherché dans la loi pour le renseignement. Cela soulève de graves problèmes d'éthique car il peut y avoir un gouffre entre un sentiment et une action. L'intérêt pour la détonique signifie-t-il qu'on souhaite faire exploser une bombe ?

B4 - Le stockage des données recueillies

Les progrès majeurs ont été faits ces derniers temps, non dans les outils d'analyse mais dans les

11 *L'étude d'impact de la loi relative au renseignement* parle de « rechercher des objectifs enfouis dans les données » mais aussi de « détecter des signaux de faible intensité ». L'étude a été éditée le 18 mars 2015 et est disponible en ligne.

outils de stockage. La DGSE a certes une expérience dans le stockage des données de communication (*French echelon*) mais c'est récemment que les grands acteurs du web comme Google, Amazon, Twitter ont changé la donne.
Ils ont trouvé le moyen de traiter en temps réel les données en les répartissant sur un grand nombre de serveurs.
Ils ont donc fait le choix de favoriser le temps de réponse de leurs logiciels plutôt que l'unicité des données, puisque les mises à jour ne peuvent pas être propagées en temps réel entre des données dupliquées sur un grand nombre de serveurs.

Les tables relationnelles de l'informatique de gestion sont remplacées par des agrégats de données. Cela suppose trois conséquences importantes. Il n'y a plus de schéma à priori qui suppose comment les données doivent être stockées. Techniquement on parle de **modèle noSQL**[12].

Par ailleurs, les données deviennent redondantes et perdent leur intégrité.
A titre d'exemple le système peut vendre à plusieurs reprises le même objet et il faudra donc prévoir une politique pour ce type particulier de dysfonctionnement.

Pareillement les données d'une même personne peuvent cumuler des incertitudes sur son âge, son métier, ses condamnations. Dans une saisie, une personne sera témoin d'un crime. Dans un autre enregistrement la personne sera l'auteur du crime.

Ces contraintes techniques ont pour premier effet de d*éporter la sémantique des données sur l'algorithme d'accès à ces données.*
C'est le traitement automatisé qui devra prévoir les relations possibles entre données. C'est donc une véritable « raison algorithmique » qui va prévaloir.
Ce sera à l'algorithme d'anticiper les erreurs d'intégrité ou les aberrations dans les données.

L'idéal consiste bien sûr à améliorer les données de façon qu'elles ne soient pas contradictoires entre elles. C'est un travail important qui conditionne la fiabilité des résultats finaux.

Les données peuvent par exemple être redressées en utilisant des bases de données plus fiables. On pense bien sûr à la base Cristina de la DGSI, non soumise au contrôle de la CNIL. Il existe également toutes les bases auxquelles les services pourraient accéder (base du fisc, base de la Sécurité Sociale ou de la CAF, bases économiques, base de réservation de tickets d'avion...)[13].

Pour dire les choses autrement, il faut passer par une identification de l'entité concernée (voiture, maison, personne)[14].

Afin de respecter la loi qui interdit l'identification des personnes dans un premier temps, il faudra que l'algorithme anonymise ce qui a aura été dés-anonymisé le temps de la fusion des données.

Cela devient de la haute voltige mais l'étude d'impact le prévoie explicitement[15].

12 Voir *Les bases de données NoSQL et le big data : comprendre et mettre en oeuvre* -- Rudi Bruchez

13 La CNIL dans sa délibération du 5 mars 2015 s'était désolée que « le traitement des données mis en oeuvre par les services spécialisés de renseignement soit susceptibles de contenir des données personnelles recueillies par d'autres canaux que ceux visés par le projet de loi » . L'*étude d'impact* donne la liste des fichiers administratifs auxquels peuvent accéder directement les services (p10).

14 La loi précise que les traitements automatisés doivent se faire « sans recueillir d'autres données que celles qui répondent à leurs paramètres de conception », ce qui ne signifie strictement rien.

15 *Projet de loi sur le renseignement – Etude d'impact*, disponible en ligne. Cf p 10 de la version pdf.

C Quelques questions qui restent ouvertes

Il est dans la nature du renseignement de générer du flou ou de la désinformation. De même, après lecture de la loi, on ne sait trop ce qu'il faut comprendre, ce qui nuit gravement à la lisibilité de la loi[16].
La notion de « réseau », par exemple, n'est pas définie[17].
Même le concept de terrorisme est très extensif. Depuis l'affaire de Tarnac, on ne sait plus exactement ce qu'est un terroriste[18] et encore moins ce que peut être un segment de la population proche du terrorisme et qu'il faudrait surveiller pendant quatre ans.

Voyons les questions qui peuvent se poser en partant du point de vue technique.

C1 - On ne sait pas exactement ce qui peut être légalement capté sur les réseaux

L'article L851-3 parle de « connexion susceptible de révéler une menace » mais le gouvernement précise que le captage porte sur les « caractéristiques techniques des communications assurées et sur la localisation des équipements terminaux »[19].
L'article du CPCE (Code des Postes et des Communications Electroniques) auquel il est fait référence ajoute qu'il ne s'agit en aucun cas du « contenu des *correspondances échangées* ou des *informations consultées*, sous quelque forme que ce soit, dans le cadre de ces communications ».

Dans cette formulation il s'agit bien de méta-données mais le texte a vieilli et une partie des activités sur les réseaux ne relève ni de la correspondance, ni de la consultation.

Les actes d'achat (d'amonitrate par exemple), de réservation (de place de train par exemple), d'abonnement (à un système d'alerte boursière par exemple), de déclaration (aux impôts par exemple), de récupération de fichier (par ftp par exemple), de signature de pétition, d'emprunt de livres et même l'écoute d'un certain type de musique sont à la limite de la définition juridique.
Il en est de même pour les signaux générés par les capteurs qui, légalement, permettent au moins de localiser le capteur. Il peut s'agir ici de n'importe quel mécanisme qu'il s'agisse de l'état piézo-électrique de l'eau, d'un compteur électrique intelligent ou d'une caméra de surveillance qu'on a installée quelque part. Inversement les signaux générés à destination d'un capteur font partie des données de connexions comme par exemple une puce RFID[20] dans une étiquette de vêtement.
D'une façon générale, sur les réseaux, l'exécution de programmes exige des données techniques de connexion pour lancer le programme (URL du programme, paramètres passés en entrée) et récupérer les résultats. Toutes ces données entrent dans le spectre de ce que la loi autorise à recueillir. Il peut s'agir de jeux, d'applications sur tablettes ou smartphones.
Pour les entreprises, il peut s'agir aussi de programmes exécutés sur le nuage (*cloud*)[21]. Ces

---

16 Le principe de clarté et d'intelligibilité de la loi est exigé dans les articles 4, 6 et 16 de la Déclaration de 1789

17 Voir le *Mémoire en réplique présenté par les députés signataires du recours dirigé contre la loi relative au renseignement* , disponible sur le site de l'Assemblée Nationale.

18 Dans l'affaire Tarnac, les auteurs d'un acte de vandalisme relativement anodin sont tombés sous l'inculpation de terrorisme car le procureur leur a attribué des propos appelant à la révolution violente (qu'ils auraient écrits dans un livre anonyme).

19 Article L34-1-VI du CPCE (Code des Postes et Communications Electroniques)

20 Les puces RFID sont la version électronique des code-barres. Elles peuvent être interrogées par un capteur local comme à une caisse de magasin ou dans un camion de transport de produits. Les informations collectées peuvent ensuite être envoyées sur les réseaux (mais ce n'est pas une nécessité liée à la technologie RFID)

21 Il s'agit de programmes que l'entreprise n'exécute pas chez elle mais sur un serveur distant. Le serveur peut être public. Il peut aussi être plus ou moins dans le périmètre de l'entreprise (cloud hybride ou réseau privé). On voit ici la nécessité d'une définition juridique de ce qu'est un « réseau ».

connections peuvent ainsi s'apparenter à de l'espionnage industriel.

C2 - On cherche en aveugle les données qui caractérisent l'intention terroriste

Le gouvernement a donné quelques explications sur les traitements automatisés en suggérant qu'ils sont élaborés à partir « d'éléments recueillis au cours d'enquête » et qui « permette de découvrir des modes particuliers de communication ». Cette explication nous laisse sur notre faim car cela voudrait dire qu'un groupe de terroristes qui communiquerait « de façon normale » passerait à travers les mailles de l'algorithme.
Comment donc caractériser une intention ?

Signal faible ou pattern

Imaginons ici qu'on s'intéresse aux signaux faibles, c'est à dire à des données qui sont atypiques par-rapport à tous les autres[22].

Il est vraisemblable qu'il y aura mise en œuvre d'un travail d'analyse qui consiste à faire parler les données recueillies[23]. Cette analyse peut être particulièrement compliquée, en plusieurs étapes, avec divers programmes et un processus constant d'amélioration (apprentissage automatique). Cela ne peut se faire que progressivement, par tâtonnement, essai-erreur. C'est donc une logique de *Big Data*.

Du fait même de la logique du Big Data, on n'a, à priori, aucune idée de ce que sont des « données caractérisant une menace terroriste » et à fortiori aucune idée des données « susceptibles » de caractériser cette menace.
Par exemple une manipulation financière atypique peut-elle caractériser une menace potentielle ?

Dans les débats et documents annexes à la loi, le gouvernement parle parfois de signaux faibles mais aussi d'une utilisation des données découvertes lors d'enquêtes, parfois de repérage d'un « *pattern* » de comportement.
La recherche d'un pattern de comportement est un mécanisme inverse de celui qu'on vient de décrire. Il consiste à repérer un agencement de données que l'on connaît déjà[24].

Si ces deux techniques se révèlent infructueuses dans une recherche de terroristes, les fichiers de travail générés existent bel et bien, par exemple le fichier des fraudeurs possibles. Le statut juridique de ces fichiers est douteux : peuvent-ils être gardés quatre ans ? Peuvent-ils être utilisés dans une autre des missions des services ?

Internet des personnes

La plupart des applications du *Big Data* donnent des informations *globales*, plus ou moins probables en tous cas non individualisées : le nombre de patients attendus par un hôpital dans les cinq ans, les meilleures rues pour un taxi en maraude, le type de musique qu'il faut diffuser...

Lorsqu'on passe au niveau *individuel*, l'analyse doit utiliser des informations d'une grande densité d'information. Cela nous éloigne un petit peu du *Big Data* qui opère sur des données insuffisamment renseignées mais en très grand nombre.

On peut ainsi prédire la solvabilité d'une personne ou le risque de devenir diabétique. On peut

22 Les signaux faibles sont utilisés pour prédire au plus tôt des phénomènes émergents (le vol d'un papillon par exemple qui annonce un tsunami)
23 L'expression *signaux faibles* est apparue dès la rédaction de l'étude d'impact et a été reprise par le gouvernement pendant les débats au parlement.
24 L'étude d'impact parle de « détection de certains comportements de communication »

définir les caractéristiques qui permettent de filtrer les employés, par exemple pour un casino[25].

Dans le domaine du renseignement, des outils américains comme *Xkeyscore* ont la réputation d'être extrêmement efficaces. Ce programme utilise les méta-données des courriels, l'adresse des sites visités, les formulaires remplis, l'utilisation de certains programmes (comme Tor), les activités sur les réseaux sociaux, les voyages (localisation des équipements), en gros tout ce qui est autorisé par la loi sur le renseignement.

Deux logiques à l'œuvre

Il y a ainsi deux logiques qui se complètent et qui, toutes deux, nécessitent de gros volumes d'information :

- celle orientée « personne » qui consiste à regrouper les informations concernant un individu ;
- et celle orientée « segment » qui consiste à faire apparaître tout ce qui est atypique (signal faible) ou au contraire ressemblant (pattern de comportement).

Dans le premier cas, en utilisant le gros volume du *Big Data*, on peut regrouper toutes sortes d'informations concernant un individu (non connu par son nom). Ces informations peuvent ensuite être elles-mêmes analysées pour en faire apparaître des régularités.

Dans le second cas, les traitements *Big Data* portent sur des données globales avec des enseignements globaux mais qui, tout en restant anonyme peuvent arriver au niveau du segment de population ou de l'individu.
Il existe par exemple des programmes qui permettent de repérer un fraudeur à la carte bancaire car tous les fraudeurs ont des comportements en commun.

Ces techniques présentent aussi un véritable intérêt social. Par exemple, un patient atteint d'une pathologie médicale complexe offre des traits caractéristiques qui peuvent être confrontés à de grandes masses de données dans le but d'avoir le traitement le plus adapté pour cette personne particulière.

C3 - Les algorithmes non contrôlés

Le « traitement automatisé » n'est pas encadré par la loi et est laissé à la libre appréciation de la CNCTR. Si du côté des données, il existe la CNIL qui contrôle les fichiers, rien n'existe du côté des algorithmes.

Or précisément, dans le cas du Big Data, ce sont les traitements qui *supportent la sémantique* des données beaucoup plus que les fichiers qui présentent un ensemble de données hétérogènes, non structurées et incompréhensibles sans traitement.

Prenons le cas de la délinquance. Il existe déjà de nombreux programmes qui aident à prévoir lieu où un acte délinquance va (sans doute) avoir lieu. Il existe aussi d'autres programmes qui aident les travailleurs sociaux à prévoir la probabilité de récidive d'un délinquant condamné.
On se souvient de la proposition de Mr Bockel qui visait à repérer les délinquants potentiels à partir des troubles de comportement des enfants de trois ans. On peut imaginer un programme qui implémente la recherche d'une délinquance potentielle à partir de données remontant à l'enfance. On conviendra qu'il encapsule une philosophie politique aussi dangereuse que la décision de créer un fichier ethnique.

Le problème est que s'il existe un contrôle des fichiers ethniques, il n'existe pas d'instance non administrative qui contrôle les programmes de lutte contre la délinquance ou la récidive.

---

25 Cf. *le Big Data* de Pierre Delort. Les sociétés spécialisées citées dans le livre sont des courtiers en données (Axciom, ChoicePoint), des analystes du marché musical (Big Champagne), des analystes dans la finance (Neo) etc...

Loi sur le renseignement : *the next frontier*

Pour trouver une *aiguille* dans une *botte de foin*, il faut déjà disposer de la botte. A ce niveau, on ne peut pas parler de « surveillance de masse » mais au moins de « captage de masse ».
Les traitements ultérieurement mis en œuvre permettront de qualifier plus précisément ce captage des données.

Dans la mesure même où ces données sont hétérogènes, incomplètes, volumineuses et volatiles, il faut trouver des techniques adaptées pour les analyser. C'est donc logiquement vers le *Big data* qu'on doit tourner son intérêt pour essayer de mesurer les enjeux.
Même en lisant en détail les documents officiels relatifs à la loi « renseignement », le citoyen reste dans le plus grand flou sur les intentions réelles de l'administration.

Deux choses sont pourtant certaines qui tiennent directement aux techniques disponibles.

Il faut d'abord avoir en mémoire que les traitements automatisés ne sont pas une baguette magique. Contrairement aux sciences exactes, les méthodes dont j'ai parlé plus haut établissent rarement un lien entre une variable recherchée (l'intention terroriste) et des variables indépendantes. La plupart du temps on n'aura que des corrélations ou des analyses assez grossières concernant un sous-ensemble (segment) des données captées.

En météorologie, le *Big Data* ne peut pas dire précisément le temps qu'il fera en un moment donné en un point précis. De même l'*Internet des personnes* ne pourra livrer que des enregistrements nombreux et pseudonymes.

Cela signifie donc que des personnes réelles seront connues grâce à des signatures numériques qui à tout moment pourront être transformées en identité civile.

Mais en fin de compte des données globales présentent le plus grand intérêt pour des autorités publiques.
Le *Big Data* peut être un recensement quotidien et un sondage d'opinion productible à volonté. Comme tous les sondages d'opinion, les résultats auront une marge d'incertitude. C'est d'ailleurs un des enjeux principaux : pouvoir mesurer l'incertitude.
Le *Big Data* peut aider aux choix d'investissements publics (crèche, route, barrage etc...) et à la conduite des politiques publiques. Le *Big Data* produit un déplacement inédit de la valeur[26].

Ces technologies sont des technologies d'avenir. La loi relative au renseignement n'est donc qu'un premier pas dans cette direction.

Dans le mouvement *open data*, ce sont les données publiques qui sont restituées aux citoyens. Le mouvement ici est inverse : ce sont les données privées qui sont captées.

Mais les traces que nous laissons dans notre vie numérique sont-elles encore des traces privées ? A qui appartiennent par exemple mes données médicales : à moi ? Aux services de santé qui peuvent en tirer des enseignements ? A ma mutuelle, mon docteur, mon Etat ?

On peut aussi se demander quelle *raison algorithmique* va s'appliquer sur ces données. Est-ce que le *Big Data* va générer un conseil pour m'éviter une intoxication médicale pourtant a priori

---

26 On dispose d'un bon exemple dans l'agro-alimentaire. Avec des capteurs et un bon historique des récoltes, on peut déterminer les meilleurs traitements à opérer sur chaque lopin de terre en fonction du temps, de la saison etc. La firme agro-alimentaire qui dispose de ces données ne vend plus des engrais mais des conseils qui vont optimiser la production en diminuant l'impact écologique.

rarissime ? Ou est-ce que le *Big Data* va automatiquement annuler mon permis de conduire vu mon état de santé ?