Loi sur le renseignement ou pratique de la surveillance automatisée ? | Le Net Expert Informatique

Un expert du Big Data m'a adressé ce texte. Il y expose clairement pourquoi, selon lui, la « détection automatisée de comportements suspects » prévue par la Loi Renseignement est très dangereuse. En un mot, mettre les gens dans des cases au moyen d'un algorithme forcément imparfait, ce n'est pas grave s'il ne s'agit que d'envoyer de la publicité ciblée, mais ça l'est beaucoup plus s'il s'agit d'envoyer des policiers interpeller des gens chez eux à 6 heures du matin.

Je vous livre ce texte :
« Depuis plusieurs années je travaille sur le big data appliqué au marketing en ligne. J'ai les mains dans le moteur du matin au soir, et lorsque j'ai appris quelle était la teneur du projet de loi qui devrait être voté le 5 mai prochain, je n'ai pu m'empêcher de frémir en essayant d'imaginer les usages possibles des techniques et des procédés annoncés. Voici quelques réflexions qui me sont venues sur ce dispositif qui pourrait transformer radicalement notre société. Je ne suis pas certain que nos députés aient une idée claire de la boîte de Pandore qu'ils s'apprêtent à ouvrir sur ordre de l'exécutif.
Je me souviens de l'aventure advenue il y a longtemps à l'un de mes oncles, militant fortement engagé dans une association (pacifique) classée franchement à gauche. Il avait vu un jour débarquer chez lui deux personnes des Renseignements Généraux, munies d'un gros dossier qui recensait en détail toutes ses activités. Juste histoire de lui faire comprendre qu'ils savaient qui il était, où il habitait, ce qu'il faisait – pourtant rien d'illégal – et qu'on le tenait à l'oeil. Une simple visite de courtoisie; ou peut-être peut-on appeler ça de l'intimidation? Tout ça s'est passé bien avant la généralisation d'Internet, des fichiers numériques et des téléphones portables. Aujourd'hui, le dossier n'aurait peut-être pas pu être porté sous le bras, ou plutôt si, sur une clé USB, contenant dix ou dix mille fois plus d'informations.
Je me souviens aussi, lorsque j'ai commencé à travailler sur des clusters, du choc que j'ai ressenti la première fois où nous avons tracé une carte utilisant des adresses IP de visiteurs (il est très facile d'obtenir des données géographiques assez fiables pour une adresse IP résidentielle). La carte mettait en évidence de manière saisissante des comportements liés directement à la provenance géographique. Les gens de mon quartier (on était déjà descendus à une échelle plus fine que celle d'une ville) avaient exactement les mêmes comportements que moi; je me suis vu dans la carte. Mon estime en a pris un coup, car j'étais rétrogradé en une seconde au rang de mouton.Mais j'ai réalisé, en regardant ce découpage coloré, à quel point ce nouvel outil nous offrait une puissance et une justesse d'analyse dont nous n'avions même pas rêvé.
Parmi les nombreux problèmes que posent cette loi, se trouve la pose de « boîtes noires » chez les fournisseurs d'accès et les hébergeurs, espionnant potentiellement tout le trafic Internet. Un malentendu assez fréquent est que l'on saura ce que vous faites en inspectant effectivement vos différentes activités en ligne. Qu'on cherchera *individuellement* vos traces d'activité suspecte. Et qu'il vous suffira de visiter quelques sites pour être visé par des investigations plus poussées. Et l'on se dit que l'on n'a rien à craindre, puisqu'on n'a certainement rien de commun avec les terroristes en puissance. Mais ce n'est pas comme ça que ces systèmes fonctionnent.
Pour qu'ils soient efficaces, ils ont besoin de modèles, dont l'utilisation s'apparente à des techniques de pêche au chalut. On attrape tout, on trie, et on garde ce qui est intéressant. Mais comment savoir ce qui est intéressant a priori ? Justement, on ne peut pas vraiment. Ça fonctionne en gros comme ça :

• Première phase, on collecte tout en vrac, sur beaucoup de monde, pendant un moment.
• Deuxième phase, on identifie le groupe d'individus que l'on recherche (mais pas directement, ou en tout cas pas uniquement en utilisant ces données), et on l'indique au système.
• Troisième phase, à partir des données qui ont été collectées sur les membres identifiés de ce groupe, le système fabrique un modèle, selon différentes méthodes.
• Et quatrième phase, on identifie tous les autres, éventuellement vous, qui ne font pas partie du groupe, parce qu'ils se conforment au même modèle.
• On continue à alimenter le système itérativement, on affine le modèle, et on continue.

Dans la pratique, le jugement humain intervient, mais si l'on cherche à étendre ce système, on peut laisser aux machines le soin d'en faire plus, et finalement opérer elles-mêmes le choix des marqueurs d'une activité « suspecte ». C'est à la fois un peu moins inquiétant (vous pouvez continuer sereinement vos recherches de nitrate d'ammonium en ligne si vous êtes agriculteur sans être soupçonné de vouloir fabriquer une bombe) et pire, car à mesure que la quantité de données disparate augmente, il va devenir compliqué de savoir pourquoi une personne a un score élevé dans une catégorie recherchée. Il ne s'agit pas de cases virtuelles que le système coche au fur et à mesure, mais de relations mathématiques et d'enchaînements entre des données dont le sens est éventuellement complètement obscur. Et on peut fort bien tomber dans la mauvaise case.

Dans le domaine du marketing, tomber dans la mauvaise case n'est pas dramatique : une publicité mal ciblée ou les recommandations absurdes d'un site de commerce en ligne n'ont jamais changé dramatiquement la vie de quiconque ; j'avais eu un bel exemple de ce genre sur le plus gros site d'e-commerce du monde il y a quelques années, où mes collègues et moi-même n'avions vu l'espace d'une matinée que des recommandations étonnantes, composées à 50% environ de prothèses de jambes. Bug manifeste du moteur de recommandations, dont nous avions eu toutes les peines du monde à nous extraire. Une fois que vous êtes lancé dans un tunnel, dans ce domaine, il est parfois difficile d'en sortir. Donc cette fois-là c'était plutôt amusant. Si un problème semblable advient sur des systèmes de surveillance, la personne qui atterrira d'un coup sur les radars des services de renseignement risque de trouver l'expérience moins ludique.

Mais on ne pourra pas surveiller tout le monde, se dit-on. En fait, si, on peut. Une des caractéristiques des systèmes dédiés au big data c'est la scalabilité linéaire. En termes moins techniques, ça signifie que pour doubler votre capacité de stockage ou de traitement, il suffit grosso modo de doubler le nombre machines dans le cluster. Un cluster, c'est un ensemble de machines (des centaines, des milliers ou plus) qui fonctionnent en parallèle et stockent chacune une partie des données dont vous les nourrissez en permanence. Le principe est d'assembler toutes ces données en les découpant d'abord en de multiples morceaux, traités en parallèle, chacun sur une machine. Au lieu d'un seul programme, vous avez mille programmes qui traitent chacun un morceau de données, tournant sur mille machines, comme s'il s'agissait d'un seul ordinateur gigantesque. Vous avez deux fois plus de données à stocker ? Rajoutez autant de machines et des disques durs. Vos traitements prennent trop de temps ? Rajoutez des machines. La beauté de la chose, c'est que ces systèmes ne sont pas plus durs à gérer quand vous passez de cent à dix mille machines. La même équipe peut s'en charger, la seule limite est le budget. Le système est extensible à l'infini. La capacité et le prix des disques durs aujourd'hui rendent éventuellement inutile la purge des données; on peut tout conserver à tout jamais. Ce n'est qu'une question de moyens.

Alors bien sûr, il faut des analystes (des statisticiens ou des spécialistes de l'intelligence artificielle) et des programmeurs pour créer les programmes qui vont établir des relations entre des données disparates. Mais là encore, beaucoup de choses peuvent être accomplies par des équipes réduites. Les algorithmes qui permettent de partir à la pêche dans l'océan des données sont maintenant rôdés, et il n'est point besoin de réinventer la roue à chaque nouveau problème. L'important est de poser la bonne question, le reste n'est qu'un détail d'exécution. De plus, grâce la puissance de ces architectures, on peut poser de multiples questions dans un temps raisonnable, ce qui n'a jamais été possible auparavant. On peut affiner la question posée, jusqu'à un grand niveau de détail. On peut obtenir des réponses à des questions que l'on n'a pas pensé à poser. Et plus le volume de données est important, plus la fiabilité des réponses, en général, augmente. Enfin, ces données restent accessibles sans délai et s'offrent pour toujours à de nouvelles analyses. Elles permettent de définir des modèles de plus en plus fins, auxquels sont comparées en temps réel les nouvelles données qu'ingurgite en continu le système. Elles permettent de classer, d'identifier, et souvent de prévoir.

Cela dit, et c'est là que la prétention d'empêcher les actes terroristes trouve sa limite, elles permettent de prévoir en termes de probabilités. Elles permettent de vous classer dans un groupe, pas de savoir vraiment si oui ou non vous allez effectivement faire telle ou telle chose, ni quand. A moins que vous n'ayez acheté une grande quantité du nitrate d'ammonium suscité par CB (ce qui serait franchement stupide), que vous ne fréquentiez assidument des individus connus pour leurs appels à la guerre sainte, et que vous n'ayez donné rendez-vous à vos copains par e-mail pour le feu d'artifice, le système ne va pas pouvoir dire quel jour et à quel endroit vous allez poser une bombe artisanale. A moins de disposer des données de centaines de personnes effectivement parties faire le jihad, et qu'elles ne permettent de construire un modèle fiable, ce qui reste à démontrer, il ne pourra pas non plus identifier de manière fiable le départ des prochains candidats. On baigne là dans l'illusion technologique. Ainsi, malgré les considérables moyens déployés aux Etats-Unis, il ne semble pas que la NSA ait atteint dans ce domaine des records d'efficacité. La France ferait-elle mieux ?

Donc, à quoi ça sert ? N'étant pas dans le secret des décideurs, je ne peux qu'imaginer : si j'étais au pouvoir et que j'avais ce gros jouet à disposition, je pourrais toujours avoir une longueur d'avance sur… tout ! Pour prévoir les grèves, les mouvements sociaux, l'agitation étudiante, les ZAD, les contestations diverses, les tendances pour les élections. Même pour la politique étrangère, l'intelligence économique, les possibilités sont infinies. Un outil extraordinaire, mille fois meilleur et plus riche en volume que tous les sondages et les compte-rendu des ex-RG. Les utilisateurs de big data dans le domaine du marketing le savent très bien: les gens mentent (sans le savoir, et croient donner des réponses sincères), mais leurs actions, elles, ne mentent pas.

Exemple au hasard, les « intérêts économiques essentiels de la nation » (un parmi la liste très large des objectifs de la loi). J'imagine fort bien des IMSI-catchers dans le quartier de la Défense, à l'écoute des managers discutant de contrats avec des firmes étrangères concurrentes de firmes françaises. Étant donnée la perméabilité entre les grandes entreprises et la haute fonction publique, je peine à croire qu'aucun conseil amical ne filtrera jamais des services de renseignement vers les directions de ces entreprises. Bien sûr on n'écoutera pas toutes les conversations des concurrents – ce qui demande trop de temps – mais il est déjà démontré qu'il suffit de connaître la liste de vos correspondants, la durée et la fréquence de vos appels pour savoir à peu près tout de votre activité et de vos projets. Les fameuses métadonnées, dont les partisans de la loi vantent la quasi-innocuité, suffiront pour tout leur dire sur vous. Le secret des affaires ? Obsolète. On pourrait faire un concours de pronostics sur tous les usages possibles de cette loi, vu son champ d'application tellement large. On serait sans doute encore à cent lieues de prévoir ce qui se passera exactement.

Mais il y a le contrôle par la commission, objectera-t-on. Je l'imagine cette commission, inondée de requêtes, combien par jour ? Dix, cent, mille ? Combien de temps passé sur chacune d'entre elles ? Comment prétendre qu'il s'agira d'autre chose qu'une chambre d'enregistrement ? Les moyens techniques permettront de rédiger des demandes par centaines, sans effort, à tel point que le contrôle de celles-ci ne deviendra plus qu'un processus de pure forme, sous l'avalanche continuelle. De toutes manières, qui garantira l'indépendance et la compétence des nominés ? Comment prétendre que remplacer tous les juges par une seule commission n'effectuant qu'un contrôle a posteriori, et dont le silence vaut accord, pourra garantir les droits de chacun ? Comment croire qu'un seul « expert technique » pourra valider tous les algorithmes utilisés ? Rien que ce dernier point me semble absurde. Ensuite, il y a la durée de conservation des données, qui est limitée. Techniquement, purger des données disparates est déjà un peu compliqué. Quant à purger des données dérivées des données brutes, pour de multiples raisons, c'est encore plus complexe. Il faudra que cet impératif soit au coeur du système dès le départ pour que cela ait une toute petite chance de fonctionner. Les paris sont ouverts.

L'exécutif se retrouverait donc doté d'un outil par définition opaque, surpuissant, qui lui permettrait de s'abstraire presque totalement du pouvoir judiciaire. Exécutif élu, rappelons-le, pour cinq ans. C'est très court, et c'est prendre un bien gros pari sur l'avenir que de mettre dans les mains de quelques personnages-clés une arme qui permet de contrôler aussi totalement tous les aspects de la vie des personnes. Et de les influencer, voire de les contraindre, quelle qu'en soit la raison. Mais après tout, si vous n'avez ni l'intention de vous syndiquer, ni de donner un avis controversé sur un forum, ni de tromper votre conjoint(e), ni de revendiquer quoi que ce soit, ni de critiquer qui que ce soit, en somme de ne pas faire quoi que ce soit que vous ne vouliez pas que la terre entière apprenne, qu'avez-vous à craindre ? C'est ce qu'on appelle une société de surveillance. La vie privée est un concept désormais obsolète, c'est presque inévitable. »

Voilà, maintenant que vous avez lu ce texte qui est bien plus argumenté que l'exemple caricatural que je vous avais donné, je vous invite à vous faire votre propre opinion, et à le partager autour de vous si vous jugez que cela peut être utile. N'hésitez pas à le transmettre aux députés qui, demain, voteront sur ce projet de loi !

PS : si mon ami a choisi l'anonymat, ce n'est pas par crainte de la police ou de la justice de la République, mais juste parce qu'il ne souhaite pas qu'un lien soit fait avec son employeur.

Après cette lecture, quel est votre avis ?
Cliquez et laissez-nous un commentaire…

Source : http://www.zdnet.fr/actualites/loi-renseignement-un-ami-expert-du-big-data-explique-le-danger-de-la-surveillance-automatisee-39818832.htm
Par @PierreCol