POINT DE VUE

Renseignement : le traitement massif de données est aussi dangereux qu'inefficace

Au nom de la lutte contre le terrorisme, la France adopte un dispositif de détection de profils de comportements suspects inopérant, coûteux et menaçant pour la vie privée.

Claude CASTELLUCCIA et Daniel LE MÉTAYER

Au motif de lutter contre le terrorisme, plusieurs pays ont adapté leur cadre juridique pour permettre un traitement massif de données à des fins de renseignement. La France s'est également engagée dans cette voie avec le projet de loi relatif au renseignement, adopté par l'Assemblée nationale en première lecture le 5 mai 2015 et en passe d'être discuté au Sénat à l'heure où nous écrivons. L'objectif affiché de ces dispositions est de doter les États de moyens de détecter des « signaux faibles », autrement dit des profils de comportements suspects qu'il conviendrait de repérer le plus tôt possible pour déjouer des projets terroristes. D'un point de vue strictement technique, on peut cependant s'interroger à la fois sur les risques liés à ces procédés et sur leur efficacité.

Le risque principal posé par un traitement massif de données est celui d'une atteinte tout aussi massive à la vie privée des citoyens. Les promoteurs de ces lois utilisent un double argument pour rassurer les populations : dans un premier temps ne seraient collectées « que » des métadonnées (ou des « données de connexion ») et celles-ci seraient « anonymisées ». Or des données sont considérées comme anonymes s'il n'est pas techniquement possible de retrouver les identités des personnes concernées.

On ne peut donc pas à la fois prétendre que des données sont anonymes et prévoir en cas de suspicion la levée de cet anonymat. Les données en question sont donc au mieux « pseudonymisées », ce qui implique, comme l'a rappelé la CNIL, qu'elles restent couvertes par la loi Informatique et Libertés. L'usage du mot « anonyme » est en l'espèce un contresens : l'objet est bien de collecter des informations liées à des personnes et, si leur état civil n'est pas dévoilé dans un premier temps, il pourra l'être ensuite.

Le paradoxe des faux positifs démontre que l'on risque d'identifier des centaines de milliers de « faux terroristes »

Revenons aux données de connexion. L'adresse de la page web à laquelle un internaute souhaite accéder est une donnée de connexion qui fournit souvent autant d'informations que la page elle-même. De manière générale, il est bien connu des spécialistes que les données de connexion sont parfois plus intrusives que les données elles-mêmes. Par exemple, le fait de savoir qu'un appel téléphonique est destiné à un cardiologue ou au bureau des alcooliques anonymes apporte plus d'informations que la conversation en soi, qui peut se réduire à une simple prise de rendez-vous. Autre illustration, l'analyse des données de géolocalisation des taxis new-yorkais a montré qu'elles pouvaient être utilisées pour inférer la religion de certains chauffeurs.

Ce n'est pas vraiment rassurant, mais est-ce efficace ? Pour cette question, un autre mot-clef apparaît de manière récurrente dans les débats, comme un remède magique : l'*algorithme* censé analyser les fameuses données de connexion pour en extraire inexorablement la liste des suspects. Il est vrai que des techniques d'apprentissage ou de classification automatique se révèlent très utiles et efficaces

dans certains scénarios, notamment dans les systèmes de recommandation ou de personnalisation à des fins commerciales. Amazon les utilise avec succès pour recommander des livres et Google pour envoyer des publicités ciblées.

Toutefois, ces algorithmes reposent sur des modèles (profils suspects dans le cas du renseignement) imparfaits qui, par conséquent, introduisent des erreurs, qu'on appelle généralement faux négatifs et faux positifs. Le « paradoxe des faux positifs », bien connu des mathématiciens, démontre que le nombre de faux positifs est considérable lorsque l'événement à identifier est rare, ce qui est le cas ici. Il faudra donc consacrer des ressources considérables pour identifier et étudier ces nombreux faux positifs qui, pour une population de 60 millions d'habitants, pourraient se chiffrer à 600 000 personnes avec un algorithme fiable à 99 % !

En d'autres termes, en faisant l'hypothèse énoncée par le gouvernement que 3 000 personnes mériteraient d'être surveillées, la probabilité qu'une personne identifiée par le système soit vraiment un terroriste serait alors de 0,5 %, ce qui est négligeable. Les ressources étant limitées, il serait plus rationnel et responsable de les concentrer sur des systèmes de surveillance ciblée plus efficaces, faute de quoi notre sécurité risque en réalité d'en pâtir. N'oublions pas que les frères Kouachi, à l'origine des actes terroristes de janvier dernier en France, étaient connus des services de renseignement : ce n'est donc pas un traitement massif de données qui aurait permis d'éviter ces drames, mais plutôt une meilleure exploitation des informations disponibles.

Une posture rationnelle reposant sur l'examen des risques et des bénéfices conduirait vraisemblablement à abandonner toutes les mesures de collecte de données non ciblées. Mais se concentrer sur des mesures ciblées ne dispenserait pas de prévoir des solutions pour en minimiser les risques. Ces solutions devraient être juridiques et organisationnelles, avec notamment des prérogatives de contrôle et des moyens significatifs pour une commission véritablement dotée de solides compétences en informatique, comme devrait l'être en France la future Commission nationale de contrôle des techniques de renseignement (CNCTR). Ce n'est pas le cas aujourd'hui, puisque la commission comporterait un seul expert. Les solutions devraient également inclure un volet technique, notamment pour assurer que les données collectées soient authentiques et intègres, ne puissent être accessibles qu'aux agents habilités et authentifiés, pour des finalités bien définies et avec une traçabilité fiable des autorisations.

Sans ces garanties techniques, tout contrôle risque de se révéler illusoire, ce qui n'est pas acceptable vu les enjeux en matière de droits individuels. En effet, le seul garde-fou possible en matière de traitement des données personnelles est l'instauration d'une véritable « responsabilité », dans le sens de l'*accountability* des Anglo-Saxons, le devoir de rendre des comptes, condition *sine qua non* de la confiance et contrepartie nécessaire de tout pouvoir. ■

© Shutterstock.com/Konstantin Yolshin

LES MÉTADONNÉES OU DONNÉES DE CONNEXION SONT BIEN PLUS BAVARDES qu'elles n'en ont l'air : savoir qu'un appel téléphonique est destiné à un cardiologue ou au bureau des alcooliques anonymes est parfois plus parlant que la conversation elle-même, qui peut se résumer à une simple prise de rendez-vous.

LES AUTEURS

Claude CASTELLUCCIA et Daniel LE MÉTAYER sont directeurs de recherche à l'institut Inria.

BIBLIOGRAPHIE

A. Berlee, **Using NYC taxi data to identify muslim taxi drivers,** billet de blog, janvier 2015. http://bit.ly/1I9p50x

J. Parra-Arnau et C. Castellucia, **Dataveillance and the false-positive paradox,** prépublication, mai 2015 : https://hal.inria.fr/hal-01157921

J. Mueller et M. G. Stewart, **Responsible counterterrorism policy,** Cato Institute, *Policy Analysis* n° 755, septembre 2014. http://bit.ly/1BHC8lq

Réagissez au Point de vue sur www.pourlascience.fr